AD________________

Award Number: DAMD17-00-1-0448

TITLE: Cox Model for Interval Censored Data in Breast Cancer Follow-up Studies

PRINCIPAL INVESTIGATOR: George Y.C. Wong, Ph.D.

CONTRACTING ORGANIZATION: Strang Cancer Prevention Center
New York, New York 10021-4601

REPORT DATE: July 2004

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	***2. REPORT DATE*** July 2004	***3. REPORT TYPE AND DATES COVERED*** Final (1 Jul 2000 - 30 Jun 2004)

4. TITLE AND SUBTITLE Cox Model for Interval Censored Data in Breast Cancer Follow-up Studies	***5. FUNDING NUMBERS*** DAMD17-00-1-0448
6. AUTHOR(S) George Y.C. Wong, Ph.D.	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Strang Cancer Prevention Center New York, New York 10021-4601 ***E-Mail:*** gwong@strang.org	***8. PERFORMING ORGANIZATION REPORT NUMBER***
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012	***10. SPONSORING / MONITORING AGENCY REPORT NUMBER***

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited	***12b. DISTRIBUTION CODE***

13. ABSTRACT (Maximum 200 Words)

The overall objective of this research proposal is semi-parametric inference of the Cox proportional hazards (PH) regression model for a survival function $\Pr(X > x \mid Z = z) = S(x \mid z) = [S_0(x)]^{e^{z\beta}}$, where X is a time-to-event variable, which is subject to interval censoring, Z represents the covariates, S_0 is a baseline survival function, and β represents the regression coefficients. The main objective of our research is to develop asymptotic inferences of the generalized maximum likelihood estimators (GMLE) of β and $S(\cdot \mid z)$. A critical limitation with GMLE under interval censoring is that it is computationally feasible only for a small data set. We therefore propose to also investigate asymptotic properties of a computationally simpler alternative to GMLE, namely two-stage estimators (TSE) of β and $S(\cdot \mid z)$ obtained by a two-stage modified Newton –Raphson algorithm involving data grouping. In the four years of our research, we have implemented a foolproof algorithm for obtaining TSE, proved consistency and established asymptotic normality for both GMLE and TSE under both discrete and continuous distributional assumptions, and proposed new diagnostic method for PH assumption. Also, we have successfully applied our asymptotic Cox regression methodology to the analysis of a large-scale, long-term breast cancer relapse follow-up study. Our results will be useful to data analysis of breast cancer relapse follow-up studies, chemoprevention intervention trials and genetic studies on familial aggregation of breast cancer and related cancers.

14. SUBJECT TERMS Breast cancer, interval-censored data, cox regression model, maximum likelihood, two-step estimation, asymptotic properties	***15. NUMBER OF PAGES*** 46
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	***18. SECURITY CLASSIFICATION OF THIS PAGE*** Unclassified	***19. SECURITY CLASSIFICATION OF ABSTRACT*** Unclassified	***20. LIMITATION OF ABSTRACT*** Unlimited

FOREWORD

Opinion, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

_X_ Where copyrighted material is quoted, permission has been obtained to use such material.

_X_ Where material from documents designated for limited distribution is quoted, permission has be obtained to use the material.

_X_ Citations of commercial organizations and trade names in this report do not constitute an official Department of the Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals", prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

N/A For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institute of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

George Wolf

A. TABLE OF CONTENTS

B. INTRODUCTION

Interval-censored (IC) data are encountered in three areas of breast cancer research. The most common application is in clinical relapse follow-up studies in which the study endpoint is disease-free survival. When a patient relapses, it is usually known that the relapse takes place between two follow-up visits, and the exact time to relapse is unknown. In statistics, we say relapse time is interval censored. Interval censoring is also encountered in breast cancer registry studies in which information on family history of cancer is updated periodically. The Strang Breast Surveillance Program for women at increased risk for breast cancer, for instance, has enlisted over 800 women with complete pedigree information which is verified and updated continuously. Family history data such as age at diagnosis of a specific cancer, or a benign but risk-conferring condition, are obtained from each registrant at each update. Time to a cancer event, and definitely time to first detection of a benign condition, are at best known to fall in the time interval between the last update and age at diagnosis. A third but increasingly important area of application of interval censoring is in breast cancer chemoprevention experiments or prevention trials, which involve the observation of one or more surrogate endpoint biomarkers (SEB) over time. The scientific questions of interest here are estimation of time for an SEB to reach a target value, and estimation of time from cessation of intake of a chemopreventive agent to the loss of its protective effect. Unfortunately, the exact values of both these time variables are known only to lie in between two successive assay inspection times. In a breast cancer follow-up study, we often encounter covariates (for instance, tumor size and nodal status in a relapse study, and baseline SEB value in a chemoprevention trial). A regression model will be needed for the analysis of such data.

Let X denote a time-to-event variable with distribution $F(x) = Pr(X \leq x)$, or equivalently, survival function $S(x) = 1 - F(x)$. In interval censoring, X is not observed and is known only to lie in an observable interval (L, R). In our previous DOD funded grant, we have made fundamental contributions to both the theory of the generalized maximum likelihood (GML) estimation of S, and the computation in connection with the inference of GML estimator (GMLE) $\hat{S}$ of S. These contributions are restricted to the case of univariate IC data without covariates.

The Cox proportional hazards regression model [2] specifies that covariates have a proportional effect on the hazard function of X. This model provides powerful means for fitting failure time observations to a distribution free model and for estimating the risk for failure associated with a vector of covariates. It is extensively used for right-censored data.

Finkelstein [3] applied the Cox model to the analysis of IC data. However, she did not establish asymptotic properties of the GMLE of the parameters in the model; moreover, she did not investigate the convergence properties of the Newton-Raphson (NR) algorithm for finding the GMLE values.

Our recent interest in IC data with covariates is driven by needs arising from two related areas of breast cancer research at Strang Cancer Prevention Center. First, we have been conducting a long-term prognostic follow-up study of breast cancer bone marrow micrometastasis (BMM) on relapse, involving 375 women. We shall need an efficient algorithm to find the GMLE of the regression coefficients of the Cox model. Second, we have just completed a one-year confirmation chemoprevention trial of of indole-3-carbinol (I3C) for breast cancer prevention. In this prevention trial we monitored the levels of two SEB's, a urinary estrogen metabolite ratio and a blood counterpart, both of which are subject to interval censoring. An earlier dose-ranging study of I3C conducted by Wong *et al* [4] has been published.

The overall aim of this research proposal is to develop statistical inference for IC data with covariates that are encountered in breast cancer relapse follow-up studies, breast cancer chemoprevention trials employing surrogate endpoint biomarkers, and in breast cancer registry follow-up studies of familial aggregation of breast and other forms of cancer. Asymptotic generalized maximum likelihood theory under the Cox regression model will be investigated and computer software package for maximum likelihood inference will be implemented.

C. BODY

C.1. Model Formulation and Likelihood Equations.

Let $Y_{K,1} < Y_{K,2} < \cdots < Y_{K,K}$ denote the follow-up times for a patient who has made K follow-up visits, in a longitudinal follow-up study. Since the number of visits for each patient may vary, K is a random positive integer. For convenience, define $Y_{K,0} = 0$ and $Y_{K,K+1} = \infty$. The time-to-event variable of interest, X, is not directly observed; instead, it is known to lie in between two successive censoring time points $(Y_{K,j}, Y_{K,j+1})$, where j =0, ..., K. Note that X is left censored if $j = 0$, strictly interval censored if $0 < j < K$, and right censored if $X > Y_{K,K}$. The observable IC data corresponding to X is given by

$$(L, R) = (Y_{K,i}, Y_{K,i+1}) \text{ if } Y_{K,i} < X \leq Y_{K,i+1},\ i = 0, 1, ..., K. \tag{1}$$

In addition to (L, R), we also observe a $p \times 1$ covariate vector Z. We assume that K and the $Y_{k,j}$'s are independent of (X, Z).

The Cox regression model for the survival function at $X = x$ given $Z = z$ is represented by

$$S(x|z) = [S_o(x)]^{e^{z\beta}}, \tag{2}$$

where $z\beta$ is the dot product of Z and β, $S_o(x)$ is a baseline survival function and β is a p-dimensional regression coefficient vector.

Let $I_i = (L_i, R_i, z_i)$, $i = 1, \ldots, n$, be a random sample of size n interval-censored observations with covariates. In terms of the original observed intervals, the likelihood function of S and b is given by

$$Ł = \prod_{i=1}^{n}((S(L_i))^{e^{bz_i}} - (S(R_i))^{e^{bz_i}}), \tag{3}$$

where S is a survival function, and b is a $p \times 1$ dimensional vector. The GMLE of (S_o, β) is a value (S, b) that maximizes (3) over all survival functions S and all $b \in \mathcal{R}^p$.

Since S_o places all probability mass on the innermost intervals of the I_i's (see Peto [5] or Turnbull [6]), it is often computationally simpler to express L in terms of innermost intervals.

We say that an interval A is an innermost interval of the I_i's if A is a nonempty finite intersection of one or more of the I_i's such that either $I_i \cap A = \emptyset$ or $I_i \cap A = A$ for each i. Suppose there are a total of m distinct innermost intervals $A_i = (\xi_i, \eta_i]$, where $\eta_i \leq \xi_{i+1}$ and $m \leq n$. Then the likelihood function (3) is equivalently given by

$$Ł = \prod_{i=1}^{n}[(\sum_{k>l_i} s_k)^{e^{z_i b}} - (\sum_{k>r_i} s_k)^{e^{z_i b}}], \tag{4}$$

where $l_i = \sup\{j : \eta_j \leq L_i\}$, $r_i = \sup\{j : \eta_j \leq R_i\}$ and $s = (s_1, \ldots, s_m)$ denote the vector of the probability weights. The log likelihood of (s, b) is

$$\mathcal{L}(s, b) = \sum_{i=1}^{n} \ln[(\sum_{k>l_i} s_k)^{e^{z_i b}} - (\sum_{k>r_i} s_k)^{e^{z_i b}}]. \tag{5}$$

Note that $(\sum_{k>r_i} s_k)^{e^{z_i b}} = 1$ if $r_i = 0$ and $(\sum_{k>l_i} s_k)^{e^{z_i b}} = 0$ if $l_i = m$.

C.2. Generalized maximum likelihood estimation.

A GMLE of (S_o, β) is a value of (s, b) that maximizes the likelihood function (5). We could follow the NR algorithm outlined by Finkelstein [3]. However, this would involve the inverse of a matrix of order $(m+p-1) \times (m+p-1)$. Since m can be potentially large when n is large, the unmodified NR algorithm is not feasible for a large data set. The major serious problem of the unmodified NR algorithm, however, is that it will almost always diverge in one iteration from any reasonable starting value. The divergence is due to the fact that the first Newton step will hit the boundary of the parameter space. Such a serious computational problem was not anticipated at the time we submitted the DOD grant.

To reduce computation burden due to dimensional effects, we propose to group the original data (L_i, R_i) and then apply a two-step modified NR algorithm to obtain the two-step estimators (TSE) of (S_o, β) based on the innermost intervals corresponding to the grouped intervals. In the **first** year of our research, we obtained a first version of a two-step modified NR algorithm to find TSE. We also carried out simulation studies to investigate sensitivity of estimated values of TSE to partition sizes. This version of two-step computational scheme, however, can diverge with some simulated IC regression data. We reported our initial findings in a preliminary manuscript in Wong and Yu [7]. In our **second** year of research, we presented an improved version of algorithm at the Era of Hope 2002 DOD Breast Cancer Research Program Meeting in Orlando (see Wong and Yu [8]).

In our **third** and **fourth** years of research, we continued to improve and update the two-step algorithm. We have tested the latest version of the algorithm on many simulated data sets and have not encountered any convergence difficulty.

In our **second** year of research, we applied our two-step estimation procedure to the Cox regression analysis of a long-term prognostic relapse follow-up study involving 375 women with unilateral T1-2N0, T1-2N1 and T3-4 breast cancer. All the patients were treated at Memorial Sloan Kettering Cancer Center and the follow-up are being conducted at Strang Cancer Prevention Center. The main objective of the study is to assess the prognostic significance of bone marrow micrometastasis (BMM) in predicting relapse. Standard clinical variables including nodal status and tumor diameter were included in the Cox model. Although we have not completely established asymptotic normality to validate the P values that were reported for the study, our two-step Cox regression analysis gave strong indication that BMM was not as predictive of relapse as previously expected (Osborne and Wong [1]). In our **second** year of research, therefore, we have moved ahead of our statement of work by making a start for Task 8. Since the BMM relapse follow-up study provides a complete and final data set that optimally satisfies our need of an empirical example to illustrate our asymptotic GML procedure for Cox regression, we have chosen to focus on this data set instead of the examples mentioned in Task 8.

Also, in the **second** year of our research, we established consistency of the GMLE of β and S_o (and hence $S(\cdot|z)$) under the following assumptions:

AS1: S_o is arbitrary and each of the censoring variables,Y_1, ..., Y_k takes on finitely many values.

AS2: S_o is arbitrary and each of the censoring variables,Y_1, ..., Y_k is continuous and some regularity conditions are imposed on either S_o or the joint distribution function G of K, Y_1, ..., Y_K.

Specifically, under AS1 and AS2

$$Pr\{\lim_{n\to\infty} \hat{\beta} = \beta\} = 1, \tag{6}$$

and

$$Pr\{\lim_{n\to\infty} \sup_{t\in H} |\hat{S}_o(t) - S_o(t)| = 0\} = 1, \tag{7}$$

where H denotes the support set of Y_1, ..., Y_K. Note that $\hat{S}_o(t)$ is guaranteed to be consistent for $t \in H$, and not elsewhere. However, the set H is not necessarily a time interval (for instance, H may be a collection of discrete points). In order for the consistency results to be more useful, we established that if S_o is continuous, and the support of Y_1, ..., Y_K is dense in $[0, T]$ for some $T > 0$, then $\hat{S}_o(t)$ is consistent for all $t \in [0, T]$. The practical implication of the denseness requirement is that pointwise consistency of $\hat{S}_o(t)$ would hold only if all the subjects in a follow-up study must be followed at very frequent close intervals.

We also established similar consistency results for the TSE, with an added assumption that the maximal length of the partition interval tends to 0 as n tends to ∞. These results are summarized in Wong and Yu [7].

Asymptotic normality is the most crucial aspect of our research because it is needed in making confidence statements and in performing hypothesis testing. In the **third** year of our research, we investigated asymptotic normality under assumptions

AS3. S_o is arbitrary and satisfies a monotonicity condition, and each of $Y_{K,1}$, ..., $Y_{K,K}$ takes on finitely many values;

AS4. S_o is as in AS3, and each of $Y_{K,1}$, ..., $Y_{K,K}$ takes on countably many values;

AS5. S_o is as in AS3, each of $Y_{K,1}$, ..., $Y_{K,K}$ is continuous and some regularity conditions are imposed on either S_o or G.

Asymptotic normality of GMLE or TSE is straightforward to establish under the finite assumption AS3. As for AS4 and AS5, we carried out extensive simulation studies to guide our research. The studies suggest that both GMLE and TSE of β and S_o are asymptotically normal under AS4. However, only GMLE and TSE of β can be asymptotically normal under AS5. We have just completed theoretical proofs to substantiate our numerical studies. Our simulation studies suggest that under AS5 asymptotic inference for GMLE and TSE of S_o, and hence $S(\cdot|z)$ will have to be accomplished via a bootstrap method. In our **fourth** and **final** year of research, we have investigated a bootstrap approach for asymptotic interval estimation of S_o when G is continuous. We have completed a manuscript that summarizes our asymptotic normality findings (see Yu and Wong [9]).

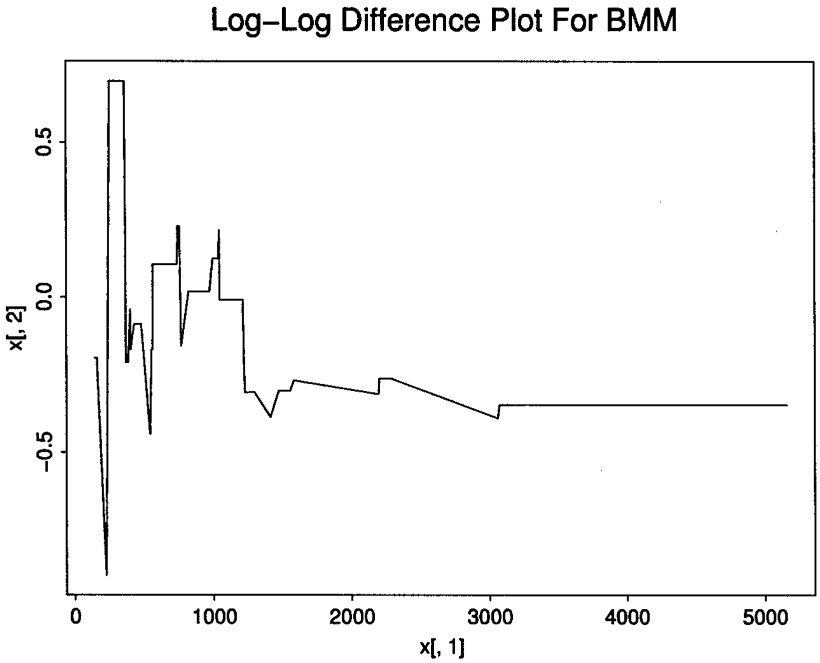


Figure 1.

Cox regression is appropriate only if proportional hazards (PH) assumption is satisfied by the data. Under the PH assumption, the log-rank test is most powerful. At present, a statistically useful diagnostic plot for PH assumption is lacking. Moreover, a formal significant test is not available. In the **third** year of our research, we provided statistical solutions to satisfy both these needs. For the diagnostic plot, we proposed to plot $\ln\hat{S}_1(t) - \ln\hat{S}_2(t)$ verse t for any two groups, where $\hat{S}$ refers to GMLE of S under interval censoring. A horizontal line should be expected if PH assumption holds. For a test for PH assumption, we proposed an asymptotic chi-square test. In our **fourth** and **final** year of research, we have completed a manuscript to report on our diagnostic solutions (see Wong and Yu [10]). We applied the diagnostic procedure to the BMM data. Figure 1 is the log difference plot for BMM+ and BMM- groups It is clear that PH assumption was inappropriate for the

BMM data. The asymptotic chi-square test gave a P-value of 0.013 indicating a significant departure from PH assumption.

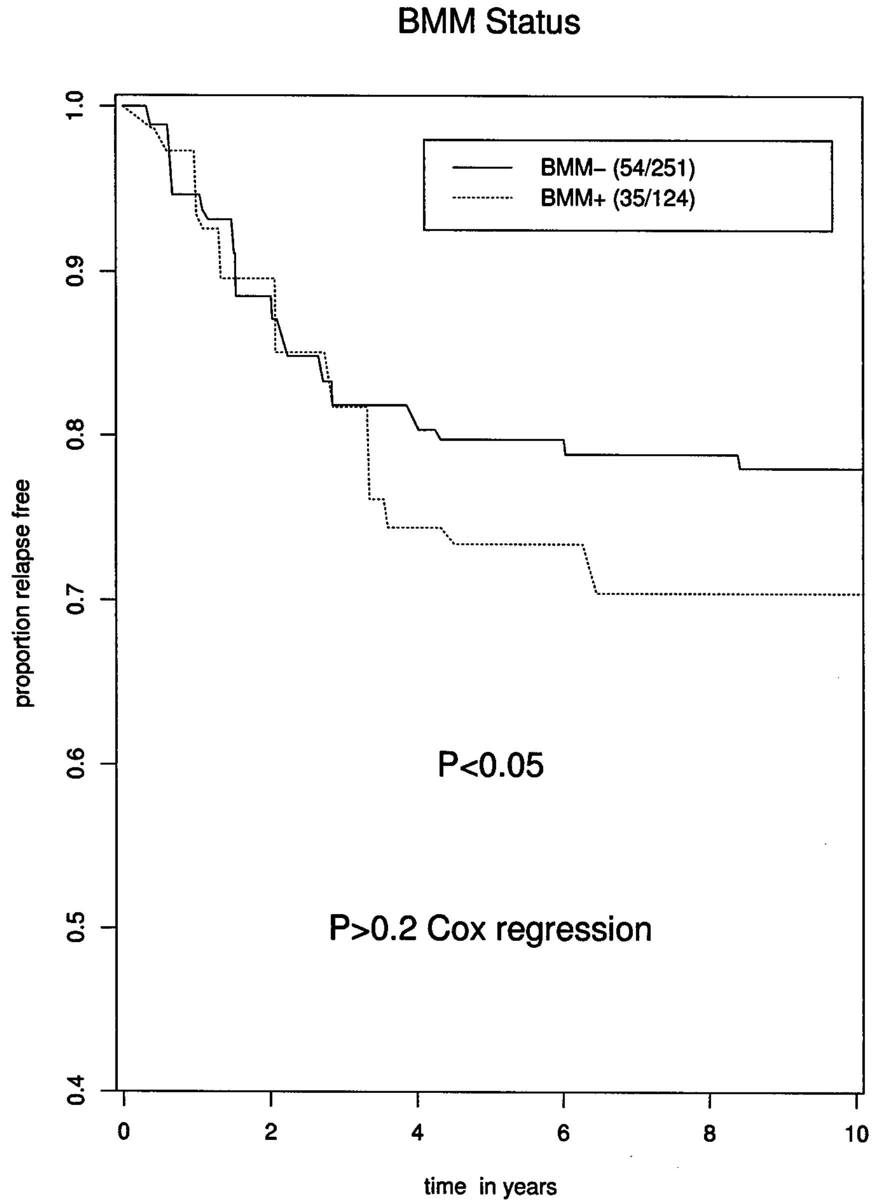


Figure 2.

Since Cox model is not appropriate for the BMM relapse follow-up study, we have little choice but to broaden our research on regression models for IC data by including semi-parametric models that do not require PH assumption to hold. In the **second** year of our research, we proposed to fit the BMM data using a linear regression model with an unknown nonparametric distribution function for the error term. We successfully derived an non-iterative algorithm to obtain the GMLE of the regression coefficient in case of a simple linear regression model. When we applied such a model to the BMM data in a univariate analysis, BMM was significant at $P < 0.05$. Incidentally, BMM was also significant by a weighted Kaplan-Meier statistics (see Pepe and Smith [11]) which does not require PH assumption. Figure 2 gives the IC-version of the Kaplan-Meier plots of the two BMM groups. This example clearly demonstrates that inappropriate application of the Cox regression model can potentially lead to an errorneous statistical conclusion. We published a statistical paper on the theory of the linear regression approach in the **third** year of our research (see Yu and Wong [12]) and gave a presentation of the novel analysis of the BMM data in the **fourth** and **final** year of our research (see Wong, Osborne and Yu [13]).

D. KEY RESEARCH ACCOMPLISHMENTS

- We have implemented a statistical algorithm for computing GMLE of the regression coefficients β and the baseline survival function S_o.
- We have implemented a statistical algorithm for computing TSE of β and S_o.
- Computer programs for both GMLE and TSE calculations have been made available to the public via the internet.
- We have proved consistency of GMLE and TSE of β and S_o under both discrete and continuous assumptions about the censoring distribution G.
- We have performed extensive simulation studies to investigate the asymptotic properties of GMLE and TSE of β and S_o. Our results have provided strong evidence that S_o is **NOT** asymptotic normal when G is continuous.
- We have derived the asymptotic normal means and covariance matrices of GMLE and TSE of β.
- When G is finite or countably infinite, we have derived the asymptotic means and covariance matrices of GMLE and TSE of S_o.
- We have proposed a diagnostic plot for checking proportional hazards assumption for Cox regression and constructed a chi-square test for assessing this assumption.
- We have completed regression analysis of a long-term breast cancer follow-up study assessing the prognostic significance of bone marrow micrometastasis in predicting relapse in a cohort of 375 women, using asymptotic GML Cox regression, weighted Kaplan-Meier statistic, semi-parametric linear regression methods.

E. REPORTABLE OUTCOMES

- One oral presentation of an abstract at 2002 ASCO Meeting ([1]).
- Two poster presentations at 2002 Era of Hope DOD Breast Cancer Research Program Meeting and 2004 Annual San Antonio Breast Cancer Symposium. ([8],[13]).
- Three published abstracts on the proposed two-stage modified Newton-Raphson algorithm and applications of the methodology to a breast cancer relapse follow-up study. ([1],[8],[13]).
- A manuscript on computation of GMLE and TSE of Cox regression parameters ([7]).
- A manuscript on consistency and asymptotic normality of GMLE and TSE ([9]).
- One published statistical paper on regression analysis of IC data ([12]).
- A manuscript on assessing the appropriateness of proportional hazards assumption for Cox regression ([10]).
- Computer programs for calculating GMLE and TSE made available for the public via the internet site *http://www.math.binghamton.edu/qyu/index.html.*

F. CONCLUSIONS

In the four years of our DOD grant, we have successfully accomplished our research objectives in developing asymptotic generalized maximum likelihood inference of Cox proportional hazards regression model with IC data. We have developed statistical algorithms that can efficiently compute GMLE and TSE of the regression coefficients β and the baseline survival function S_o for any reasonable sample size. We have proved consistency of GMLE and TSE of β and S_o under both discrete and continuous assumptions about the interval censoring distribution G. We have established asymptotic normality for GMLE and TSE of β for G unrestricted. When G is continuous, however, we have numerically demonstrated that GMLE and TSE of S_o are not asymptotically normal. In the **fourth** and **final** year of our DOD grant, we have investigated a bootstrap method for the asymptotic interval estimates of S_o.

Cox regression is appropriate only if proportional hazards (PH) assumption is satisfied by the data. We have proposed a useful diagnostic plot for PH assumption and validated a chi-square test for it.

In our **fourth** and **final** year of research, we have completed the final version of a computer software for asymptotic confidence intervals and hypothesis testing for GMLE and TSE of β and $S(\cdot|z)$. We have also completed the data analysis of the BMM prognostic study.

The results which we have established will be useful to clinicians analyzing breast cancer relapse follow-up prognostic studies, breast cancer researchers pursuing chemoprevention intervention trials involving surrogate endpoints biomarkers, and genetic epidemiologists conducting studies on familial aggregation of breast cancer and related cancers.

G. REFERENCES

[1] Osborne, MP and Wong GYC. (2002). Breast cancer bone marrow micrometastases: a long-term prognostic study of systemic tumor cell burden on relapse. *Proceedings of American Society of Clinical Oncology*, 21, #228.

[2] Cox, D.R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. B,* 34 187-220.

[3] Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 42 845-854.

[4] Wong, GY, Bradlow, HL, Sepkovic, D, Mehl, S, Mailman, J, and Osborne, MP (1997). A dose-ranging study of indole-3-carbinol for breast cancer prevention. *Journal of Cellular Biochemistry Supplement* 28/29 111-116.

[5] Peto, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.* 22, 86-91.

[6] Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser.* B, 38, 290-295.

[7] Wong, G.Y.C. and Yu, Q.Q. Estimation under the Cox regression model with interval-censored data. (Manuscript for submission).

[8] Wong, G.Y.C and Yu, Q.Q (2002). A two-step Newton-Raphson algorithm for generalized maximum likelihood estimation of Cox regression model for interval-censored data in breast cancer follow-up studies. *Proceedings of the Era of Hope 2002 DOD Breast Cancer Research Program Meeting*, P50-6.

[9] Yu, Q.Q and Wong, G.Y.C. Asymptotic properties of the GMLE and TSE under the Cox regression model with interval-censored data. (Under preparation).

[10] Wong, G.Y.C. and Yu, Q.Q. A Test for checking Cox's model with a dichotomous covariate. (Manuscript for submission).

[11] Pepe, M.S. and Fleming T.R. (1991). Weighted Kaplan-Meier Statistics: Large sample and optimality considerations. *J.R.S.S. B* 53, 341-352.

[12] Yu, Q.Q. and Wong, G.Y.C. (2003). Semi-parametric MLE in simple linear regression with interval-censored data. *Communications in Statistics-Simulation and Computation*, 32 147-164.

[13] Wong, G.Y.C, Osborne, M.P. and Yu, Q.Q. (2003). Bone marrow micrometastases is a significant predictor of long-term relapse-free survival for breast cancer by a non-proportional hazards model. *Proceedings of the 26th Annual San Antonio Breast Cancer Symposium*, #420.

Reference (1)

228 **General Poster, Sat, 1:00 PM - 5:00 PM**

Breast cancer bone marrow micrometastases: a long-term prognostic study of systemic tumor cell burden on relapse. *M. P. Osborne, G. Wong; Cornell Univ Medcl College, New York, NY*

The presence of bone marrow micrometastases (BMM) detected at the time of initial surgery has been shown to predict relapse-free survival (RFS) by different investigators. We have conducted a long-term (BMM) study involving 375 women with unilateral T1–2N0 (56%), T1–2N1 (42%), and T3–4 (2%) breast cancer. BMM was determined using monoclonal antibodies to cytokertatin. Median follow-up was 8 years (range 1 month -15 years). BMM was detected in 124 (35%) patients. Logistic regression analysis did not show any correlation between BMM prevalence and standard prognostic indicators, including lymph node status and tumor diameter. The number of BMM cells detected, representing the systemic tumor cell burden (TCB), was also examine for its relationship with standard prognostic variables using lognormal regression analysis. No significant correlation was established. Recent methodology of interval-censored survival analysis showed that BMM prevalence does not predict relapse. At a median follow-up of 2.5 years, TCB was a significant univariate predictor of early relapse. In a multivariate analysis, lymph node positive patients with high TCB had a significantly shorter RFS than all other patients. However, at a median follow-up of 8 years, the prognostic significant of TCB in node positive patients was reduced. Nevertheless, the combination of a positive nodal status and a high TCB still identifies a prognostically poor subgroup with a 5-year RFS around 50%. There were only 22 (6%) such patients in our study, most of whom had received adjuvant therapy. This group may reflect early stage IV disease; therefore, this group deserves close attention in a larger study to verify the poor prognosis we have observed, as well as to evaluate new treatment protocols to improve RFS.

230 **General Poster, Sat, 1:00 PM - 5:00 PM**

Docetaxel + epirubicin and docetaxel + doxorubicin are effective and well tolerated first-line treatments for metastatic breast cancer. *R. C. F. Leonard, K. M. Malinovszky, P. J. Barrett-Lee, A. Howell, S. R. Johnston; Southwest Wales Cancer Inst, Swansea Wales, UK; Velindre Hospital, Cardiff, UK; Christie Hospital, Manchester, UK; Royal Marsden Hospital, London, UK*

Background and objectives: Docetaxel (Taxotere®) + anthracyclines have been shown to be efficacious and tolerable first-line chemotherapies for metastatic breast cancer. A large phase III study, TAX 306, showed that docetaxel + doxorubicin is more effective than the old standard doxorubicin + cyclophosphamide. The aim of this on-going prospective study is to evaluate the efficacy and safety of docetaxel + doxorubicin or epirubicin in a UK, community-based, real-life setting and to compare the results with those obtained from TAX 306. This is an interim report. Methods: Patients received docetaxel 75 mg/m^2 and either doxorubicin 50 mg/m^2 or epirubicin 75 mg/m^2 D1 q3 weeks, at the treating clinician's discretion. To date, 225 patients (WHO performance status 0–2) have been enrolled, of which 79 received doxorubicin and 146 epirubicin. The recommended anthracycline dose was administered in 94% of cycles for doxorubicin and 90% for epirubicin. All patients were evaluated for safety and 158 patients (70%), who received ≥2 chemotherapy cycles, were evaluated for tumor response. Results: Ninety-three (59%) patients had a response (partial or complete) to either regimen; however, there were no significant differences in response between the two treatment groups (doxorubicin 61%, epirubicin 58%; p=0.71). The response rate is similar to that achieved with docetaxel-doxorubicin in TAX 306 (59%). Neutropenia was the most common adverse event, with 91 patients (40%) requiring hospitalization; 64 (28%) with neutropenia and 38 (17%) with febrile neutropenia or neutropenic sepsis (some patients had both). Overall, there was no significant difference in frequency of hospitalization between the two arms (doxorubicin 53%, epirubicin 41%; p=0.08). There was one death from neutropenic sepsis (0.4%) in the epirubicin arm. Conclusion: The interim results from this real-life study confirm previous findings from a large phase III study: docetaxel + doxorubicin or epirubicin are effective and well tolerated first-line treatments for metastatic breast cancer. Both regimens exhibited similar efficacy and safety.

229 **General Poster, Sat, 1:00 PM - 5:(**

Reduction of metastases in breast cancer - patients treated with preope hormone replacement therapy (HRT): a retrospective analysis in 972 w *F. Schuetz, I. J. Diel, T.von Holst, U. Haus, G. Bastert; Univers Heidelberg, Dept Gynaecology and Obstetrics, Heidelberg, Ger IMEREM, Nuernberg, Germany*

Substitution of estrogenes and progestins is the most common therap prophylaxis for postmenopausal discomforts like hot flushes, osteopo etc. However in the majority of studies long term HRT has been assoc with an slightly increased risk of breast cancer. On the other hand pa with preoperative HRT have a lower mortality and a longer overall-sur For further investigation we examined 972 patients between 45 an years at the time of the first diagnosis of breast cancer with and wi HRT with regard to the incidence of bone metastases. 241 patients premenopausal (mean 48.0±3.0y), 731 were postmenop (55.5±4.4y), 303 of them received HRT (group HRT+) and 428 pa not (group HRT-). Patients of group HRT+ received estrogenes o minimum of 1 year (mean 5.5±4.0y). Although the tumor size of HRT- was significantly higher than in group HRT+ (5.5±1.8 vs. 2.1± nodal status, S-phase fraction, grading and hormone-receptor s showed no differences. Adjuvant treatment in the postmenopausal g were also not significantly different. In regard to the incidence of met ses patients without HRT have significantly ($p<0.001$) more bone met ses (49 patients of group HRT- versus 5 patients of group HRT+). pulmonal (18:2) and liver (28:6) metastases were significantly frequent in patients without an preoperative HRT. It was shown *in viv* in clinical bisphosphonate trials that a normalization of bone metaboli able to reduce subsequent bone metastases efficiently. We may as that the incidence of bone metastases can be reduced by normalizing metabolism (soil) and lowering conditions of tumor cell seeding by HR

231 **General Poster, Sat, 1:00 PM - 5:0**

Weekly docetaxel with or without corticosteroid premedication as fi second-line treatment in patients (pts) with metastatic breast cancer (MBC *Stemmler, W. Mair, M. Stauch, J. Papke, G. Deutsch, W. Abenhard Dorn, C. Kentenich, C. Jackisch, S. Leinung, O. Brudler, B. Vehling-Ka J. Stamp, M. Malekmohammadi, V. Heinemann; Medical Dept III, Un sity of Munich, Munich, Germany; Oncologic Practice, Munich, Germ Deaconess Hospital, Karlsruhe, Germany; Academic Teaching Hos Aschaffenburg, Germany; Dept of Gynecology, University of Mu Munich, Germany; Dept of Gynecology, University of Münster, Mün Germany; Aventis Pharma Deutschland, Bad Soden, Germany*

Objective: This large phase II study was designed to evaluate (1) efficacy of a weekly schedule of docetaxel as first or second-line the and (2) toxicity with or without corticosteroid premedication. Methods: pts (median age 58, range 37–80) with MBC were included in the Docetaxel was given at weekly doses of 35 mg/m^2 x6, followed by a 2- rest. Additional cycles with 3 weeks of treatment and 2 weeks of rest administered until disease progression. The first 34 pts were randomiz receive dexamethasone 8 mg prior to docetaxel or no premedica Results: To date, 110 pts (first-line 16, second-line 94) were evaluab toxicity, all had measurable disease and ECOG performance status ≤2 99 pts who received >6 doses of docetaxel were evaluable for respons total of 1367 doses of docetaxel were given (median 10, range 1- Response (first/second-line): CR 1/9 pts (7.1%/10.6%), PR 4/28 (28.6%/32.9%), SD 4/28 pts (28.6%/32.9%) and PD 5/20 pts (35. 23.6%). Overall response rate was 35.7%/43.5%. Median time to prog sion was 6.6/5.8 months and median survival was 14.2 months (not reached for first-line pts). Hematologic toxicity was usually mild moderate, with no difference between the two premedication gro Non-hematologic toxicity in percent of pts (+/- steroids) included: Gra and II pleural effusions 1.1%/5.9%, edema 2.2%/5.9%, lacrimal 10.8%/17.6%, epistaxis 9.7%/17.6%, nail changes 12.9%/35.3%. Tr ment was delayed due to neutropenia in 99 cycles (7.2%), and omitte 67 cycles (4.9%). Dose reductions of level 1/2 (-5/-10 mg/m^2/dose) w required in 10/5 pts (9%/4.5%). Conclusions: The results of this st confirm that a weekly schedule of docetaxel 35 mg/m^2 is efficient and s Corticosteroid premedication as generally recommended is mandatory.

Reference [7]

ESTIMATION UNDER THE COX REGRESSION MODEL WITH INTERVAL-CENSORED DATA

By George Y. C. Wong [1] and Qiqing Yu [1]

Strang Cancer Prevention Center, 428 E 72nd Street, New York, NY 10021, USA
and
Department of Mathematical Sciences, SUNY, Binghamton, NY 13902, USA
email address: qyu@math.binghamton.edu

Current version: 8/30/2001
AMS 1991 subject classification: Primary 62 G07; Secondary 62 J99.

Key words and phrases: Interval-censored data, Cox regression, grouped data, algorithms.

Abstract: We consider the estimation problem under the Cox proportional hazards model with interval - censored data. Under this model, the survival function S at time x given the covariate z satisfies $S(x|z) = (S_0(x))^{e^{\beta' z}}$, where S_0 is a baseline survival function. β and S_0 are estimated by the generalized maximum likelihood estimator (GMLE). The Newton-Raphson (NR) method and the profile likelihood (PL) method for obtaining the GMLE do not work most of the time in our simulation study and our cancer research data, as the maximum value of the likelihood is achieved outside the parameter space and the GMLE is achieved on the boundary of the parameter space in these cases. We propose a different algorithm to compute the GMLE. The algorithm is able to search for the GMLE along the boundary as well as within the parameter space. We also propose to group the data to reduce the dimension of the parameter space. Simulation results suggest that the estimator is consistent. We apply our method to the cancer cosmesis study and to another cancer research data.

1. Introduction We consider the estimation problem under the Cox proportional hazards model (Cox, 1972) with interval -censored data.

Interval-censored data are encountered in many areas of the medical research. For instance, in clinical cancer relapse follow-up studies, the study endpoint is disease-free survival. When a patient relapses, it is usually known that the relapse takes place between two follow-up visits, and the exact time to relapse is unknown. Let X denote a time-to-event variable with distribution $F(x) = Pr(X \leq x)$, or equivalently, survival function $S(x) = 1 - F(x)$. Then X is not observed and is only known to lie in an observable interval $(L, R]$. A standard method is to use the generalized maximum likelihood estimator (GMLE) to estimate S.

The Cox proportional hazards model specifies that covariates have a proportional effect on the hazard function of the failure time distribution, namely, the survival function at $X = x$ given $Z = z$ can be represented by

$$S(x|z) = [S_0(x)]^{e^{z\beta}}, \tag{1.1}$$

where $z\beta = z'\beta$, *i.e.*, the dot product of two vectors, $S_0(x)$ is a baseline survival function and β is a p dimensional regression coefficient vector. This model provides powerful means for fitting failure time observations to a distribution free model and for estimating the risk for failure associated with a vector of covariates.

Finkelstein (1986) applied the Cox model to the analysis of interval-censored data. Huang (1996) studied the asymptotic properties of the GMLE of the regression parameters in the Cox model with current status data, which is a special case of interval-censored data. There is no explicit expression for the GMLE, thus one has to use numerical methods. Finkelstein suggested to use the Newton-Raphson (NR) method to compute the GMLE.

[1] Partially supported by Army Grant DAMD17-99-1-9390 and DAMD17-00-1-0448.

Huang suggested to use a profile likelihood (PL) approach. However, both authors did not really compute any estimates under Cox's model in their paper, except under the restricted model that $\beta = 0$. In fact, in our simulation studies these methods do not work most of the time. Also these methods do not work in the data set used in Finkelstein (1986). We shall illustrate the reasons via a simple data example. To-date, how to find the GMLE in the Cox regression model remains unsettled.

There are three computational problems in the Cox regression model approach with interval-censored data:
(a) a value of the likelihood function at a non-monotone estimate of S_0 can be greater than that at the GMLE;
(b) the GMLE may occur on the boundary of the parameter space;
(c) there are often too many parameters to be estimated.

If case (b) holds, one may apply the NR method to the boundary of the parameter space. However, this approach is not feasible if n is moderate (see Appendix 2).

In this paper, we propose a different algorithm to deal with Troubles (a) and (b). The main idea is to allow searching the GMLE along the boundary, as well as within the parameter space, dynamically. We also propose to group the data in a certain manner to reduce the dimension of the parameter space. In a cancer research data set, the number of parameters is reduced from 59 to 8. The paper is organized as follows. In Section 2, we introduce notations and the model. In Section 3, we introduce our procedure. In Section 4, we apply the method to two cancer research data. In Section 5, we present simulation results. The results suggest that the new proposed procedure is feasible and the estimates of parameters and their variances are consistent. Some tedious calculation is put in Appendix 1. In Appendix 2, we present a simple example to illustrate that the NR method, a scaled NR method and a PL method do not work.

2. The Cox regression model and notations. We shall first describe the model. Let $Y_{K,1} < Y_{K,2} < \cdots < Y_{K,K}$ denote the follow-up times for a patient who made K follow-up visits, in a longitudinal follow-up study. Since the number of visits for each patient may vary, K is a positive random integer. For convenience, define $Y_{K,0} = 0$ and $Y_{K,K+1} = \infty$. The time-to-event variable of interest, X, is not directly observed; instead, it is known to lie in between two successive censoring time points $(Y_{K,j}, Y_{K,j+1})$, where j =0, ..., K. Note that X is left censored if $j = 0$, strictly interval censored if $0 < j < K$, and right censored if $X > Y_{K,K}$. The observable interval-censored data corresponding to X is given by

$$(L,R) = (Y_{K,i}, Y_{K,i+1}) \text{ if } Y_{K,i} < X \leq Y_{K,i+1},\ i = 0, 1, ..., K. \tag{2.1}$$

In addition to (L, R), we also observe a $p \times 1$ covariate vector Z. We assume that K and $Y_{k,j}$'s are independent of (X, Z).

Let (L_i, R_i, z_i), i = 1, ..., n, be a random sample of interval-censored observations with covariates. The likelihood function is

$$\mathrm{L} = \prod_{i=1}^{n}((S(L_i))^{e^{bz_i}} - (S(R_i))^{e^{bz_i}}), \tag{2.2}$$

where S is a survival function, and b is a $p \times 1$ dimensional vector. The GMLE of (S_0, β) is a value of (S, b) that maximizes the likelihood function L in (2.2) over all possible survival functions S and $b \in \mathcal{R}^p$.

To compute the GMLE, we shall introduce some notations. We say a set is a finite intersection of observed intervals I_i's (see (2.1)) with end-points L_i and R_i if the set is an intersection of one or more observed intervals. We say an interval A is an innermost interval if it is a nonempty finite intersection of the observed intervals such that for each observed interval I_i, I_i and A are either disjoint or nested. Suppose there are totally m innermost intervals. Let ξ_i and η_i, $i = 1, 2, \ldots, m$ denote the left and right end-points of the ith innermost intervals, where $\eta_i \leq \xi_{i+1}$. For convenience, define $\eta_0 = -\infty$.

It follows from Peto (1973) or Turnbull (1976) that the GMLE of S places all probability weights on the innermost intervals. Thus it suffices to maximizes

$$\mathrm{L} = \prod_{i=1}^{n}((S_{l_i})^{e^{bz_i}} - (S_{r_i})^{e^{bz_i}}) \text{ or } \mathcal{L} = \sum_{i=1}^{n}\log((S_{l_i})^{e^{bz_i}} - (S_{r_i})^{e^{bz_i}}),$$

where $S_i = S(\eta_i)$, $l_i = \sup\{j :\ \eta_j \leq L_i, j \geq 0\}$ and $r_i = \sup\{j :\ \eta_j \leq R_i, j \geq 0\}$, One can construct a numerical example that $-\mathcal{L}$ is not convex and has multiple stationary points, even subject to the monotone condition.

3. Methods. We shall introduce the new procedure in this section.

3.1. Grouping. Grouping can reduce the dimension of the parameter space. We group the original data as follows. Partition the whole range of data points into p subintervals. For example, group the data in the unit of month, half-year or year. Let $t_1 < \cdots < t_p$ be the partition points. Denote data after grouping by (L_i^*, R_i^*), where $L_i^* = t_j$ if $t_j \leq L_i < L_{j+1}$ and $R_i^* = R_k$ if $R_{j-1} < R_i \leq R_j$, $t_0 = -\infty$ and $t_{p+1} = \infty$.

After grouping, compute the GMLE using the method in §3.2 based on the grouped data (L_i^*, R_i^*)'s. We shall assume that the data hereafter have been grouped and abusing notation, we denote them by (L_i, R_i)'s instead.

3.2. A feasible algorithm for the GMLE. Abusing notations, we identify S with a vector $(S_1, ..., S_m)$. Similarly, we identify $S^{(i)}$ with $(S_1^{(i)}, ..., S_m^{(i)})$.

Step 0. Let $b^{(0)} = 0$ be the initial estimate of β and the GMLE of a survival function with observations (L_j, R_j), $j = 1, ..., n$ be the initial estimate of $S^{(0)}$.

Step $i+1$ $(i \geq 0)$. Let $b^{(i)}$ and $S^{(i)}$ be the updated values of b and S at Step i. Do b-step and S-step as follows.

* (**b-step**) With $S = S^{(i)}$ fixed, find a b so that the likelihood function $\mathcal{L}(S^{(i)}, \cdot)$ increases. Denote the up-dated estimate b by $b^{(i+1)}$. In particular, one can use the NR method to obtain the maximum point b of the likelihood function with the given $S = S^{(i)}$.
* (**S-step**) With $b = b^{(i+1)}$ fixed, search a non-increasing S so that the likelihood function $\mathcal{L}(\cdot, b^{(i+1)})$ is maximized (or increases). Denote the up-dated estimate S by $S^{(i+1)}$. In order to guarantee the up-dated S_0 is nondecreasing, proceed as follows. Let $S^{(i+1),0} = S^{(i)}$. At Sub-step j $(j = 1, ..., m)$, update $(S_1, ..., S_m)$ by $(S_1^{(i+1),j}, ...S_m^{(i+1),j})$, where $S_h^{(i+1),j} = S_{j,u_o}$ and $S_{j,u} = \begin{cases} \frac{S_h^{(i+1),j-1}+u}{1+u} & \text{if } h < j, \\ \frac{S_h^{(i+1),j-1}}{1+u} & \text{if } h \geq j, \end{cases}$ $h = 1, ..., m$, $u_o > 0$ is a number maximizing $L(b^{(i+1)}, S_{\cdot,u})$ where $S_{\cdot,u} = (S_{1,u}, ..., S_{m,u})$.

  Note: If such u_o is difficult to choose, one may choose a u_o satisfying

$$L(b^{(i+1)}, S^{(i+1),j}) > L(b^{(i+1)}, S^{(i+1),j-1}). \tag{3.1}$$

  In particular, if $\frac{\partial}{\partial u} \ln L(b^{(i+1)}, S_{\cdot,u})\big|_{u=0} > 0$, $u_o = c^k \frac{\partial}{\partial u} \ln L(b^{(i+1)}, S_{\cdot,u})\big|_{u=0}$, where $S_{\cdot,u} = (S_{1,u}, ..., S_{m,u})$ and k is the smallest non-negative integer that is smaller than K_o such that Inequality (3.1) holds.

Stop at convergence.

Expression of the partial derivatives can be found in Appendix 1.

Remark. Let p_i be the weight on the innermost interval $(\xi_i, \eta_i]$, $p = (p_1, ..., p_m)$ and $p^{(i)}$ the updated value of p at the ith step. Since $S(\eta_i) = p_{i+1} + \cdots + p_m$, the S-step can also be replace by the p-step as follows.

* (**p-step**) With $b = b^{(i+1)}$ fixed, search a non-increasing S so that the likelihood function $\mathcal{L}(\cdot, b^{(i+1)})$ is maximized (or increases). Let $p^{(i+1),0} = p^{(i)}$. At Sub-step j $(j = 1, ..., m)$, update $(p_1, ..., p_m)$ by $(p_1^{(i+1),j}, ...p_m^{(i+1),j})$, where $p_h^{(i+1),j} = p_{j,u_o}$ and $p_{j,u} = \begin{cases} \frac{p_h^{(i+1),j-1}+u}{1+u} & \text{if } h = j, \\ \frac{p_h^{(i+1),j-1}}{1+u} & \text{if } h \neq j, \end{cases}$ $h = 1, ..., m$, $u_o > 0$ is a number maximizing $L(b^{(i)}, S_{\cdot,u})$ where $S_{\cdot,u} = (S_{1,u}, ..., S_{m,u})$ and $S_{i,u} = p_{i+1,u} + \cdots + p_{m,u}$.

  Moreover, the restriction $u > 0$ can be replaced by $u > -p_h^{(i+1),j-1}$.

3.3. Estimation of the covariance matrix. It is well known that the GMLE of a survival function based on interval-censored observation (L_i, R_i)'s may not be strictly decreasing on the set $\{\eta_i : i = 0, 1, ..., m\}$. Thus the GMLE $\hat{S}$ of S_0 may satisfies $\hat{S}(\eta_i) = \hat{S}(\eta_{i+1})$ for some i. In the latter case, according to our simulation results, it is often that the empirical Fisher information matrix is ill-conditioned or singular, unless we modify $\mathcal{L}$ as follows.

Delete the innermost intervals at which the GMLE of S_0 assigns no weight. Let $(\xi_i^*, \eta_i^*]$, $i = 1, ..., m^*$, be the remaining innermost intervals. Modify $\mathcal{L}$ as

$$\mathcal{L}^* = \sum_{i=1}^{n} \log\{[S_{l_i^*}]^{e^{bz_i}} - [S_{r_i^*}]^{e^{bz_i}}\},$$

where $l_i^* = \sup\{j : \eta_j^* \leq L_i\}$ and $r_i^* = \sup\{j : \eta_j^* \leq R_i\}$. Then use the inverse of the empirical Fisher information matrix corresponding to this modification as the estimate of the covariance matrix.

This modification works well in our applications and simulations.

4. Application.

Example 4.1. We apply our procedure to a breast cancer relapse follow-up study based on data obtained from 375 women with stages I - III unilateral invasive breast cancer surgically treated at Memorial Sloan-Kettering Cancer Center between 1985 and 2001. The median follow-up duration was 7.4 years. Relapse time was given by the time interval between surgery and the initial relapse. A relapse that took place between two successive follow-up visits was regarded as interval censored. If a patient did not relapse toward the end of the study, then her relapse time was right censored. Of the 375 observations, 288 were right censored (no relapse), 20 were left censored and 67 were strictly interval censored (87 relapses). The tumor diameter (i.e., the diameter of the tumor which was removed in surgery), the number of lymph nodes with metastisis, and bone marrow micrometastasis (BMM) were determined for each woman at the time of surgery. Denote the three covariates by U_1, U_2 and U_3, respectively. The first variable, (tumor size), is discretized as $Z_1 = \mathbf{1}_{(X_1 \geq 3)}$; the second variable, (number of lymph nodes), is discretized as $Z_2 = \mathbf{1}_{(X_2 \geq 1)}$; and the BMM variable is discretized as $Z_3 = \begin{cases} 1 & \text{if the number of metastasis bone marrow cells} > 14, \\ 0 & \text{otherwise.} \end{cases}$

Table 1. Results on estimating β with cancer data.

data		β_1	β_2	β_3
original		0.580 (0.229)	0.886 (0.242)	0.344 (0.293)
grouped in months		0.576 (0.229)	0.888 (0.242)	0.338 (0.293)
grouped in half-years		0.577 (0.229)	0.884 (0.242)	0.337 (0.293)
grouped in years		0.569 (0.229)	0.889 (0.242)	0.328 (0.293)

The NR and PL method do not work for this data set due the phenomenon (a). We use the proposed procedure to obtain the GMLE. The GMLE's of the regression coefficients based on the original data and the grouped data are presented in Table 1. We only present the GMLE of the regression coefficients.

The number of parameters to be estimated is 56 in the original data. There are 54 innermost intervals with 21 of them having positive weights. We also grouped data by months, half-years and years. The numbers of parameters are 16, 12 and 8, respectively. Thus properly grouping indeed reduces the dimension of the parameter space. The GMLE's of β_i's and their standard errors are presented in Table 1 too. SE's in Table 1 were computed by the procedure introduced in §3.3.

Example 4.2. (Breast cancer cosmesis study). The data set can be found in Finkelstein and Wolfe (1985). We refer the reader to that paper for a complete description of the study. Finkelstein (1986) applies the Cox regression model to compare the patients who received adjuvant chemotheraphy to those who did not in this study. So there is one covariate, the indicator that the patient received adjuvant chemotheraphy. There are 94 patients in the study. There are 30 innermost intervals. That is, $m = 30$ and $p = 1$. Thus there are 29 parameters related to the underline survival function and one parameter related to covariate in the Cox model. The GMLE of β is 0.80 with a standard error 0.29. A GMLE of S_o is a step function taking jumps at 13 points as given below.

$$\begin{pmatrix} t: & 5 & 7 & 8 & 12 & 17 & 19 & 20 & 25 & 31 & 34 & 39 & 48 \\ \hat{S}: & 0.97 & 0.96 & 0.92 & 0.87 & 0.83 & 0.81 & 0.70 & 0.65 & 0.58 & 0.57 & 0.43 & 0.27 \end{pmatrix}$$

Note we only give 12 points above, as there are two points that are very close. One can check that case (a) is true for this data set, and the NR or PL method does not work.

5. Simulation. In order to assess the asymptotic properties of the GMLE, we carried out simulation studies. Hereafter denote $Exp(\mu, \sigma)$ a distribution with the pdf $f(x) = \frac{1}{\sigma}e^{-[\frac{x-\mu}{\sigma}+1]}\mathbf{1}_{(x>\mu-\sigma)}$. The underlying distributions are as follows: X has an exponential distribution $Exp(5,5)$. The covariate $Z = (Z_1, Z_2, Z_3)'$,, where Z_1, Z_2 and Z_3 are i.i.d. from a discrete distribution with pdf $f(i) = \frac{i}{\sum_{j=1}^{6} j}$, $i = 1, ..., 6$. (L, R) is generated by the following scheme:

$$(L,R) = \begin{cases} (0,U) & \text{if } X \leq U, \\ (20,\infty) & \text{if } X > 20, \\ (U+kV, U+(k+1)V) & \text{if } X \leq 20,\ kV < X-U \leq (k+1)V \text{ and } k \geq 1. \end{cases}$$

where $U \sim U(0,2)$ and $V \sim U(0,2.3)$.

We carried out simulation with sample sizes 50, 200 and 400, and with 1000 replications for each case. The sample means and the standard errors (SE) are presented in Tables 2, 3 and 4. In grouping data, we tried lengths of intervals 3, 5, and 8.

Table 2. Grouping effect on estimating β when $n = 200$.

grouping width		β_1	β_2	β_3
	true value	-0.1	0.2	-0.1
0	GMLE	-0.091 (0.032)	0.194 (0.034)	-0.091 (0.033)
3	GMLE	-0.100 (0.041)	0.211 (0.044)	-0.100 (0.040)
5	GMLE	-0.105 (0.046)	0.214 (0.048)	-0.104 (0.047)
8	GMLE	-0.109 (0.055)	0.213 (0.057)	-0.109 (0.053)

Table 3. Simulation results on convergence.

	β_1	β_2	β_3
$n = 50$	-0.123 (0.142)	0.238 (0.127)	-0.092 (0.129)
$n = 200$	-0.109 (0.055)	0.213 (0.057)	-0.109 (0.053)
$n = 400$	-0.105 (0.037)	0.207 (0.039)	-0.107 (0.038)
true value	-0.1	0.2	-0.1

Table 4. Simulation results on estimating of variances when $n = 400$

β_i	$\overline{\hat{\beta}_i}$	SE of $\hat{\beta}_i$	$\overline{\hat{\sigma}_{\hat{\beta}_i}}$	SE of $\hat{\sigma}_{\hat{\beta}_i}$
width =8	n=400			
-0.1	-0.105	0.037	0.037	0.002
0.2	0.207	0.039	0.038	0.002
-0.1	-0.107	0.038	0.037	0.002
width =5	n=400			
	-0.105610	0.033031	0.033080	0.001190
	0.206348	0.034651	0.034211	0.001282
	-0.105921	0.033598	0.033039	0.001098
width =8	n=200			
	-0.108608	0.055201	0.042939	0.003206
	0.213256	0.056661	0.047957	0.004768
	-0.108610	0.053448	0.042963	0.003164
width =5	n=200			
width =3	n=200			
width =0	n=200			
width =8	n=50			

Table 2 indicates that the GMLE does not change much after grouping, though the SE increases, as expected. Table 3 presents simulation results based on grouping intervals of length 8. The table suggests that the GMLE of the regression coefficients based on grouped data are consistent. Table 4 present the mean of $\hat{\beta}_i$, the SE of $\hat{\beta}_i$, the mean of the estimates of the SE of $\hat{\beta}_i$ and the SE of the estimates of the SE of $\hat{\beta}_i$. Table 4 suggests that estimates of the variances of the GMLE's match the sample variances quite well.

Appendix 1.

We derive the partial derivatives needed in §3 here. Given the right boundary points of the IM's, say, $t_0 = -\infty < t_1 < \cdots < t_m = \infty$, denote $S_i = S(t_i)$, $S_{l_i} = S(L_i)$ and $S_{r_i} = S(R_i)$, *i.e.*, l_i (resp. r_i) represents the index j such that $S_j = S(L_i)$ (resp. R_i). Then

$$\ln((S_{l_i})^{e^{x_i b}} - (S_{r_i})^{e^{x_i b}}) = \begin{cases} \ln(1 - (S_{r_i})^{e^{x_i b}}) & \text{if } l_i = 0, \\ e^{x_i b} \ln S_{l_i} & \text{if } r_i = m, \\ \ln((S_{l_i})^{e^{x_i b}} - (S_{r_i})^{e^{x_i b}}) & \text{if } 0 < l_i < r_i < m. \end{cases}$$

$$\frac{\partial \ln L}{\partial b} = \sum_{i=1}^{n} e^{z_i b} z_i D_i, \text{ where } D_i = \begin{cases} \frac{\mathcal{B}_0(l_i) - \mathcal{B}_0(r_i)}{(S_{l_i})^{e^{z_i b}} - (S_{r_i})^{e^{z_i b}}} & \text{if } r_i \neq m \\ \ln S_{l_i} & \text{if } r_i = m, \end{cases}$$

$$\text{and } \mathcal{B}_0(h_i) = \begin{cases} (S_{h_i})^{e^{z_i b}} \ln(S_{h_i}) & \text{if } 0 < h_i < m, \\ 0 & \text{if } h_i = 0. \end{cases}$$

$$\frac{\partial^2 \ln L}{\partial b \partial b'} = \sum_{i=1}^{n} z_i z_i' \left\{ e^{z_i b} D_i - \mathbf{1}_{(r_i \neq m)} e^{2 z_i b} \left[D_i^2 - \frac{\mathcal{B}_1(l_i) - \mathcal{B}_1(r_i)}{(S_{l_i})^{e^{z_i b}} - (S_{r_i})^{e^{z_i b}}} \right] \right\},$$

where $\mathcal{B}_1(h_i) = \begin{cases} (S_{h_i})^{e^{z_i b}} (\ln(S_{h_i}))^2 & \text{if } 0 < h_i < m, \\ 0 & \text{if } h_i = 0. \end{cases}.$

Write

$$U_{jk}(u) = \begin{cases} \frac{S_j}{1+u} & \text{if } k \leq j, \\ \frac{S_j + u}{1+u} & \text{if } k > j. \end{cases} = \begin{cases} \frac{S_j}{1+u} & \text{if } k \leq j, \\ \frac{S_j - 1}{1+u} + 1 & \text{if } k > j. \end{cases}$$

Then

$$\frac{\partial U_{jk}(u)}{\partial u} = \begin{cases} -\frac{S_j}{(1+u)^2} & \text{if } k \leq j, \\ -\frac{S_j - 1}{(1+u)^2} & \text{if } k > j. \end{cases}$$

$$\frac{\partial^2 U_{jk}(u)}{\partial u^2} = \begin{cases} 2\frac{S_j}{(1+u)^3} & \text{if } k \leq j, \\ 2\frac{S_j - 1}{(1+u)^3} & \text{if } k > j. \end{cases}$$

Moreover, $U_{ik}(0) = S_j$,

$$\frac{\partial U_{jk}(0)}{\partial u} = \begin{cases} -S_j & \text{if } k \leq j \text{ and } 0 < k < m, \\ 1 - S_j & \text{if } k > j \text{ and } 0 < k < m, \\ 0 & \text{otherwise.} \end{cases}$$

$$\frac{\partial^2 U_{jk}(0)}{\partial u^2} = \begin{cases} 2S_j & \text{if } k \leq j \text{ and } 0 < k < m, \\ 2(S_j - 1) & \text{if } k > j \text{ and } 0 < k < m, \\ 0 & \text{otherwise.} \end{cases}$$

Abusing notations, write $S_j = S_j(u) = U_{jk}(u)$.

$$\frac{\partial \ln L}{\partial u}\Big|_{u=0} = \sum_{i=1}^{n} e^{z_i b} \left\{ \mathbf{1}_{(r_i \neq m)} \left[\frac{(S_{l_i})^{e^{z_i b} - 1} U'_{l_i k}(u) - (S_{r_i})^{e^{z_i b} - 1} U'_{r_i k}(u)}{(S_{l_i})^{e^{z_i b}} - (S_{r_i})^{e^{z_i b}}} \right] + \mathbf{1}_{(r_i = m)} \frac{U'_{l_i k}(u)}{S_{l_i}} \right\}.$$

Appendix 2

We use a simple numerical example to illustrate why the various existing algorithms do not work for the GMLE.

§1. Consider fitting Cox's regression model with five observations (L_i, R_i, Z_i): (2,5,0), (3,4,0), (5,9,1), (1,6,1), (7,8,0). It can be viewed as data from two groups, corresponding to $Z_i = 0$ or 1. Then, the innermost intervals are (3,4), (5,6) and (7,8). Let the weights on these innermost intervals be p_1, p_2 and p_3, with $p_1 + p_2 + p_3 = 1$ and $p_i \geq 0$. Note that the baseline survival function S satisfies $S(4-) = 1$, $S(4) = S(6-) = p_2 + p_3$, $S(6) = S(8-) = p_3$ and $S(8) = 0$. For this example, it is more convenient to express the likelihood as a function of p_i's rather than S. The likelihood is $L = p_1^2 p_3 (1 - p_3^{e^\beta})(p_2 + p_3)^{e^\beta}$. Since $p_1 + p_2 + p_3 = 1$, in view of L, it is simpler to write the log likelihood as

$$l = \log[p_1^2 p_3 (1 - p_1)^{e^\beta} (1 - p_3^{e^\beta})]. \tag{A.1}$$

The parameter space is $\Omega = \{(\beta, p_1, p_3) : \beta \in (-\infty, \infty), p_1 \geq 0, p_3 \geq 0, p + 1 + p_3 \leq 1\}$ with $p_2 = 1 - p_1 - p_3$. For convenience, we write $\alpha = e^\beta$ hereafter. Thus,

$$l = 2\log p_1 + \log p_3 + \alpha \log(1 - p_1) + \log(1 - p_3^\alpha).$$

Since the likelihood function has only three variables, it can be shown by direct derivation that the GMLE of (β, p_1, p_2, p_3) is approximately $(-0.461, 2/3, 0, 1/3)$.

In general, the likelihood is not so simple and one needs to compute the GMLE by numerical methods. We shall illustrate by this example that several naive numerical methods fail to yield the GMLE. They include: (a) the Newton-Raphson (NR) method; (b) the scaled NR method and (c) the profile likelihood (PL) method. Finally, we shall illustrate by this example why our new algorithm can yield the GMLE. The main difference is that the first three algorithms cannot search the GMLE along the line $p_2 = 0$ (or $p_1 + p_3 = 1$), while the new algorithm can. Note that the GMLE is on boundary $p_2 = 0$.

§2. In order to apply the NR method, we need to compute the partial derivatives.

$$\frac{\partial l}{\partial \alpha} = \log(1-p_1) - \frac{p_3^\alpha \log p_3}{1-p_3^\alpha},\ \frac{\partial^2 l}{\partial \alpha^2} = -\frac{p_3^\alpha (\log p_3)^2}{(1-p_3^\alpha)^2}, \tag{A.2}$$

$$\frac{\partial l}{\partial p_1} = \frac{2}{p_1} - \frac{\alpha}{1-p_1} \qquad \text{and } \frac{\partial l}{\partial p_3} = \frac{1}{p_3} - \frac{\alpha p_3^{\alpha-1}}{1-p_3^\alpha}. \tag{A.3}$$

§2.1. (The NR method). At the GMLE $(p_1, p_3) = (2/3, 1/3)$ with $\beta = -0.461$, Equation (A.3) yields that the gradient in (p_1, p_3) is $(1.11, 1.11)$. In other words, as (p_1, p_2) moves towards outside the parameter space, the likelihood increases. Thus the maximum value of L without the restriction of the parameter space can only be achieved outside the parameter space. The NR yields the unrestricted maximum point of L. Thus the solution to the NR algorithm is not the GMLE.

§2.2. A scaled NR method is as follows.

Let $\beta = 0$ or $\alpha = 1$ be the initial value, and let the GMLE (or SCE) of (p_1, p_3) at $\beta = 0$ be the initial value of (p_1, p_3).

Step 1. Maximize L over β with given up-dated (p_1, p_3) using the NR method.

Step 2. Maximize L over (p_1, p_3) with up-dated β using a scaled NR method, that is, scale the increments $\triangle p_i$'s in the original NR algorithm by a constant c so that the updated (p_1, p_3) remains in the parameter space.

Repeat Steps 1 and 2 until convergence.

However, it does not work in this example. In particular, in the initial step. we have $\beta = 0$ (or $\alpha = 1$) and $(p_1, p_3) = (3/5, 2/5)$. In Step 1, L is maximized by $\alpha = -\frac{\log 2}{\log 0.4} \approx 0.76$ (see Eq. (A.4)). In Step 2, by Equation (A.3), the gradient at $(p_1, p_3) = (3/5, 2/5)$ is $(1.44, 0.61)$. Thus (p_1, p_3) should be up-dated to $(\frac{3}{5}+1.44x, \frac{2}{5}+0.61x)$ for some $x \geq 0$. If $x > 0$, it violates the constraint $p_1 + p_3 \leq 1$. Thus the algorithm stops at $S(4) = p_2 + p_3 = 2/5$ and $S(6) = p_3 = 2/5$ with $\beta = \log 0.76$ $(= -0.274)$, which is not the GMLE.

§2.3. A PL approach is as follows:

The initial step and Step 1 are the same as in the scaled NR method above.

Step 2 (p_1-substep). Maximize L over p_1 with up-dated p_3 and β.

Step 3 (p_3-substep). Maximize L over p_3 with up-dated p_1 and β.

Repeat Steps 1, 2 and 3 until convergence.

However, the PL method still does not work. In particular, at Step 1, $\alpha = 0.76$, $p_1 = 0.6$ and $p_3 = 0.4$. The gradient at $(p_1, p_3) = (0.6, 0.4)$ is $(1.44, 0.61)$. Thus we move (p_1, p_3) either to $(0.6 + 1.44x, 0.4)$ with $x \geq 0$ (p_1 substep), or to $(0.6, 0.4 + 0.61x)$ with $x \geq 0$ (p_3 substep). If $x > 0$, both the p_1-substep and the p_3-subtep will move (p_1, p_3) outside the parameter space. Consequently, it will stop at the value which is not the GMLE.

§2.4. There are three line segments in the boundary of the parameter space in (p_1, p_3). They are $p_1 = 0$, $p_3 = 0$ and $p_1 + p_3 = 1$. One can find the value that maximizes the likelihood on these line segments separately, using the NR method, and they check which is the GMLE. This approach works in this example. However, if there are m p_i's, we need to consider the subsets of the boundary corresponding to one $p_i = 0$, two $p_i = 0$, ..., $m - 2$ $p_i = 0$. Thus the order is $O(m^{m/2})$. When m is large, this approach is not feasible.

§3. We now illustrate why the new algorithm works. Our new algorithm is as follows.

The initial step. Let the GMLE of (p_1, p_2, p_3) be the initial value of the (p_1, p_2, p_3) and $\alpha = 1$ the initial value of α.

β-step. Maximize L over β with up-dated p_i's.

S-step. Each S-step consists of 3 substeps: p_1-substep, p_2-substep, p_3-substep.

p_1-substep. Consider a transformation $p_{11}(u) = \frac{p_1+u}{1+u}$, $p_{12}(u) = \frac{p_2}{1+u}$, and $p_{13}(u) = \frac{p_3}{1+u}$, $u > 0$. This transformation ensures that $(p_{11}(u), p_{12}(u), p_{13}(u))$ remains in the parameter space of (p_1, p_2, p_3) for each $u > 0$. Let u_o be the value of u that maximizes $L(\beta, p_{11}(u), p_{12}(u), p_{13}(u))$ over $u \geq 0$, with β and p_i's given in the previous step. Then up-date p_i by $p_i = p_{1i}(u_o)$, $i = 1$, 2, 3.

p_2-substep. Consider another transformation $p_{21}(u) = \frac{p_1}{1+u}$, $p_{22}(u) = \frac{p_2+u}{1+u}$, and $p_{23}(u) = \frac{p_3}{1+u}$. If $\frac{\partial}{\partial u}\ln L(\beta, p_{21}(u), p_{23}(u))\big|_{u=0} > 0$, choose a $u_o > 0$ that maximizes $L(\beta, p_{21}(u), p_{22}(u), p_{23}(u))$ over $u \geq 0$, with β and p_i's given in the previous step. Up-date p_i by $p_i = p_{2i}(u_o)$, $i = 1, 2, 3$.

p_3-substep. Consider a further new transformation $p_{31}(u) = \frac{p_1}{1+u}$, $p_{32}(u) = \frac{p_2}{1+u}$, and $p_{33}(u) = \frac{p_3+u}{1+u}$. If $\frac{\partial}{\partial u}\ln L(\beta, p_{31}(u), p_{33}(u))\big|_{u=0} > 0$, choose a $u_o > 0$ that maximizes $L(\beta, p_{31}(u), p_{32}(u), p_{33}(u))$ over $u \geq 0$, with β and p_i's given in the previous step. Up-date p_i by $p_i = p_{3i}(u_o)$, $i = 1, 2, 3$.

At the p_1-substep of the initial iteration step, by Eq. (A.5), $\frac{\partial}{\partial u}\ln L(\beta, p_{11}(u), p_{13}(u))\big|_{u=0} = 0.51 > 0$ at $(p_1, p_3) = (0.6, 0.4)$, and $u_o \approx 0.1$ maximizes $L(\beta, p_{21}(u), p_{23}(u))$. At this step (p_1, p_2, p_3) is up-dated to $(\frac{0.7}{1.1}, 0, \frac{0.4}{1.1})$ $(= (0.636, 0.364))$. At the p_2-substep and p_3-substep, by Equations (A.6) and (A.7), $\frac{\partial}{\partial u}\ln L(\beta, p_{i1}(u), p_{i3}(u))\big|_{u=0} < 0$, $i = 2, 3$, thus no change is made. However, since (p_1, p_2, p_3) is changed at this S-step, β (or α) will also be change at the next β-step.

In fact, in the next β-step, β is up-dated to $\ln 0.69 = -0.371$. In the p_1-substep, $\frac{\partial}{\partial u}\ln L(\beta, p_{11}(u), p_{13}(u))\big|_{u=0} = 0.14 > 0$, L is maximized by $(p_1, p_3) = (\frac{0.742}{1.142}, \frac{0.4}{1.142}) = (0.65, 0.35)$ with $u_o = 0.042$. by Equations (A.6) and (A.7), $\frac{\partial}{\partial u}\ln L(\beta, p_{i1}(u), p_{i3}(u))\big|_{u=0} < 0$, $i = 2, 3$, thus no change is made. However, since (p_1, p_2, p_3) is changed at this S-step, β (or α) will also be change at the next β-step.

Iteratively repeat these two steps, the algorithm will yield the GMLE $(\beta, p_1, p_2, p_3) = (-0.461, 2/3, 0, 1/3)$.

Remark 2. Recall that $p_i^{(0)}$ is the GMLE of p_i when $\beta = 0$. Let $\hat{p}_i$ be the GMLE under Cox's model. According to our observation, it is often the case that if $p_i^{(0)} = 0$ then $\hat{p}_i = 0$ too. It is not clear that whether it is indeed true that

$$p_i^{(0)} = 0 \text{ iff } \hat{p}_i = 0.$$

If this is true, then one can delete the p_i's for which $p_i^{(0)} = 0$ in the algorithm to reduce the dimension of the parameter space. Moreover, after this elimination, the NR method will work too, since the GMLE is in the interior of the parameter space. However, both the sufficient and the necessary conditions may not hold.

§4. The following is the details of deriving the GMLE directly. There are only 3 variables in L, by direct examination, one can find that the maximum value of L is outside the parameter space and the GMLE of (p_1, p_2, p_3) is on the boundary of the parameter space. Moreover, the GMLE of (p_1, p_2, p_3) is on the subspace $p_2 = 0$, as $L = 0$ if $p_1 = 0$ or $p_3 = 0$. If $p_2 = 0$, $L = (1-p_3)^2 p_3^{1+\alpha}(1-p_3^\alpha)$. $\frac{\partial l}{\partial \alpha} = \log p_3 - \frac{p_3^\alpha \log p_3}{1-p_3^\alpha} = \log p_3 \frac{1-2p_3^\alpha}{1-p_3^\alpha} = 0$ implies that unless $p_3 = 1$, we have $p_3^\alpha = 1/2$ or $p_3 = 2^{-\alpha}$ or

$$\alpha = -\log 2/\log p_3. \tag{A.4}$$

For each fixed p_3, if $\alpha = -\log 2/\log p_3$, L achieves its maximum $(1-p_3)^2 p_3^{1-\log 2/\log p_3}(1-p_3)^{-\log 2/\log p_3}$. The GMLE can be found by plotting the graph of (p_3, L).

§5. In this section, we shall derive the partial derivatives needed in §3.

$$\frac{\partial}{\partial u}p_{11}(u) = \frac{1-p_1}{(1+u)^2},\ \frac{\partial}{\partial u}p_{12}(u) = \frac{-p_2}{(1+u)^2},\ \frac{\partial}{\partial u}p_{11}(u) = \frac{-p_3}{(1+u)^2}.$$

$$\frac{\partial}{\partial u}\ln L(\beta, p_{11}(u), p_{13}(u))\big|_{u=0} = \frac{2(1-p_1)}{p_1} + \frac{\alpha p_3^\alpha}{1-p_3^\alpha} - \alpha - 1 \tag{A.5}$$

$$\frac{\partial}{\partial u}p_{21}(u) = \frac{-p_1}{(1+u)^2},\ \frac{\partial}{\partial u}p_{22}(u) = \frac{1-p_2}{(1+u)^2},\ \frac{\partial}{\partial u}p_{23}(u) = \frac{-p_3}{(1+u)^2}.$$

$$\frac{\partial}{\partial u}\ln L(\beta, p_{21}(u), p_{23}(u))\big|_{u=0} = -2 - 1 + p_3\frac{\alpha p_3^{\alpha-1}}{1-p_3^\alpha} + \alpha\frac{p_1}{1-p_1}. \tag{A.6}$$

$$\frac{\partial}{\partial u}p_{31}(u) = \frac{-p_1}{(1+u)^2},\ \frac{\partial}{\partial u}p_{32}(u) = \frac{-p_2}{(1+u)^2},\ \frac{\partial}{\partial u}p_{33}(u) = \frac{1-p_3}{(1+u)^2}.$$

$$\frac{\partial}{\partial u}\ln L(\beta, p_{31}(u), p_{33}(u))\big|_{u=0} = -2 + \frac{1-p_3}{p_3} - \frac{\alpha p_3^{\alpha-1}(1-p_3)}{1-p_3^\alpha} + \alpha\frac{p_1}{1-p_1} \tag{A.7}$$

References.

* Peto, R. (1973). Experimental survival curve for interval-censored data. *Applied Statistics.* 22, 86-91.
* Cox, D.R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. B,* 34, 187-220.
* Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics,* 42, 845-854.
* Huang, J. (1996). Efficient estimation for proportional hazards models with interval censoring. *Ann. Statist.,* 24, 540-568.
* Groeneboom, P. and Wellner, J.A. (1992). Information bounds and nonparametric maximum likelihood estimation. *Birkhäuser Verlag, Basel.*
* Schick, A. and Yu, Q. Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scan. J. of Statist.* 27, 45-55.
* Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser.* B. 38, 290-295.
* Yu, Q. Q., Schick, A., Li, L. X. and Wong, G. Y. C. (1998). Asymptotic properties of the GMLE of a survival function with case 2 interval-censored data. *Statist. & Prob. Let.* 37, 223-228.
* Yu, Q. Q., Schick, A., Li, L. X. and Wong, G. Y. C. (1998). Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times. *Canadian Journal of Statistics* 26, 619-627.

Reference (8)

A TWO-STEP NEWTON-RAPHSON ALGORITHM FOR GENERALIZED MAXIMUM LIKELIHOOD ESTIMATION OF COX REGRESSION MODEL FOR INTERVAL-CENSORED DATA IN BREAST CANCER FOLLOW-UP STUDIES

George Y. C. Wong, Ph.D., and Qiqing Yu, Ph.D.

Strang Cancer Prevention Center, 428 East 72nd Street,
New York, NY 10021; and Department of Mathematical
Sciences, State University of New York,
Binghamton, NY 13902

gwong@strang.org and qyu@math.binghamton.edu

Interval-censored data (IC) arise naturally in breast cancer follow-up studies in which the exact value of a time-to-event variable X cannot be observed but is known to lie in a time interval usually defined by two successive follow-up time points. Examples of such an X variable include time to relapse and time for the value of a biomarker to reach a target value in a breast cancer prevention trial.

This research proposal is concerned with survival analysis of X in the presence of p covariates denoted by the vector Z. Let S(x|z)=Pr(X>x|Z=z) denote the survival function of X at Z=z. Cox regression model for S(x|z) is given by $S(x|z)=[S_o(x)]^{e^{\beta z}}$, where S_o is a baseline survival function, and □ is a px1 vector of regression coefficients. Generalized maximum likelihood estimates (GMLE) of □ and S_o have to be obtained iteratively. The usual Newton-Raphson (NR) algorithm will most of the time not work owing to dimensional constraint and an inherent tendency towards convergence to local maxima in the IC situation. We propose a two-step NP algorithm to overcome these problems. At the first step, □ is updated with the usual NR algorithm. This step should present little computational difficulty because the dimension p is usually manageable. Let $x_1 < \Lambda < x_m$ denote the right-end points of the innermost intervals for the IC data. At the second step of our propose algorithm, the $S(x_i)$ are updated successively using univariate NR algorithm one component at a time, and the monotonicity constraint $S(x_1) \geq \Lambda \geq S(x_m)$ is maintained at every move.

When m is large, we propose to partition the data into groups to reduce computational burden. Monte Carlo simulations indicate that the GMLE's are quite robust against partition size.

We have applied the two-step NR algorithm successfully to a breast cancer follow-up study involving 375 patients and with m=21. The purpose of the study is to assess the prognostic significance of bone marrow micrometastasis in predicting relapse and survival.

Our two-step NR algorithm provides for the first time a feasible computational algorithm that will enable researchers to perform Cox regression analysis of IC data from breast cancer follow-up studies of reasonable dimensions.

The U.S. Army Medical Research and Materiel Command under DAMD17-00-1-0448 supported this work.

Reference [10]

A Test For Cox's Proportional Hazards Model With A Dichotomous Covariate

By George Y. C. Wong [1] and Qiqing Yu [1]

Strang Cancer Prevention Center, 428 E 72nd Street, New York, NY 10021, USA
email address: gwong@strang.org
and
Department of Mathematical Sciences, SUNY, Binghamton, NY 13902, USA
email address: qyu@math.binghamton.edu

Current version: 6/4/2002
AMS 1991 subject classification: Primary 62 G10; Secondary 62 A10.

Key words and phrases: two-sample problem, testing hypothesis, Cox regression model.

Abstract: The logrank test for the two-sample problem is very popular in medical research and is most powerful under Cox's proportional hazards model. There is a diagnostic plot for visualize whether data fit Cox's model, but there is no test available for testing whether a continuous and censored data set indeed fits the model. We propose such a test. We demonstrate by cancer research data that if our test suggests that Cox's model is appropriate, both the logrank test and the test based on the weighted Kaplan-Meier statistic give consistent results. If our test suggests that Cox's model is not appropriate, the logrank test suggests that the difference between two survival functions is not significant, while test based on the weighted Kaplan-Meier statistic indicates that the difference is significant.

1. Introduction.

Cox's proportional hazards model has often been used in comparison of two survival functions. However, there is no discussion in the literature on how to test whether the model is appropriate for a continous and censored data set. We shall address this issue in this paper.

In medical research, an important question that is asked frequently is whether one group of patients has a higher survival rate than another group of patients, or whether a new treatment is better than another treatment. This is called a two-sample problem. Let X_{ij}, $j = 1, ..., n_i$, be survival times of patients in Group i, $i = 1, 2$. X_{ij} has an unknown survival function S_i. The null hypothesis for two-sample problems is

$$H_0\text{: } S_1 = S_2 \ (S_1(x) = S_2(x) \text{ for all } x),$$

against the alternative hypothesis

$$H_1\text{: } S_1 \geq S_2 \ (S_1(x) \geq S_2(x) \text{ for all } x \text{ yet } S_1(x) > S_2(x) \text{ for some } x).$$

Another alternative is

$$H_2\text{: } S_1 \neq S_2 \ (S_1(x) \neq S_2(x) \text{ for some } x).$$

It is often that the value of X_{ij} is not exactly observed, but is only known to lie within two time points, say L_{ij} and R_{ij}. For example, relapse time of a cancer patient is only known to occur between two consecutive follow-up times, or is right censored if relapse has not taken place by the last follow-up time.

[1] Partially supported by Army Grant DAMD17-00-1-0448.

In the latter case, one can denote L_{ij} the last follow-up time and $R_{ij} = \infty$. We say such observations are interval censored. In this paper, we assume that X_{ij} may subject to interval censoring. Thus, one observes (L_{ij}, R_{ij}), where $\begin{cases} X_{ij} \in [L_{ij}, R_{ij}] & \text{if } L_{ij} = R_{ij}, \\ X_{ij} \in (L_{ij}, R_{ij}] & \text{if } L_{ij} < R_{ij}. \end{cases}$ Under right censoring, the observations are equivalent to (M_{ij}, δ_{ij}), where $M_{ij} = L_{ij}$ and δ_{ij} is the indicator function of the event $R_{ij} = \infty$.

By letting Z be the indicator function of the event that the observation is from Group 1, one can use regression models to test the two-sample problem. In particular, if data satisfy Cox's proportional hazards model, namely, $S(t|Z) = (S_o(t))^{e^{\beta Z}}$, where S_o is a baseline survival function, the null hopthesis H_0 becomes H_0': $\beta = 0$. Note that the test becomes the logrank test when data are right censored. The logrank test is most powerful under the Cox regression model (see Miller (1981)).

In order to utilize the Cox regression model, it is important to test whether the Cox regression model is appropriate for a given data set. Since

$$\log(-\log(S_1(x))) = e^{\beta} + \log(-\log(S_o(x))$$

and

$$\log(-\log(S_2(x))) = \log(-\log(S_o(x)),$$

it is suggested in several textbooks (see Cox (1994) and Lee (1992)) to use the log-log survival plot to see whether the Cox regression model is appropriate. In particular, letting $\hat{S}_i$ be the generalized maximum likelihood estimator (GMLE) of the survival function based on observations from Group i (i = 1, 2), it is suggested to plot the two curves

$$y = \log(-\log(\hat{S}_1(x))) \text{ and } y = \log(-\log(\hat{S}_2(x)))$$

on the same graph. If the two curves are roughly parallel, then the Cox model is appropriate. However, there is no test available for testing whether the two curves are parallel. We present two examples (see Examples 3.1 and 3.4) that even though the data does not fit Cox's model, it is difficult to judge whether log-log plots "appear" parallel.

Cox (1984, p.150) suggests a testing procedure for discrete data, with the test statistic $H = \hat{h}_2(t)/\hat{h}_1(t)$, where $\hat{h}_i$ is an estimate of the discrete hazard of Sample i. However, if the data are continuous, *e.g.*, there is no tie among data, this approach is not appropriate.

In Section 2, we propose an improved diagnostic plotting procedure and a test on whether the two curves are parallel with right-censored data or interval-censored data. In Section 3, we apply the new procedures to several data sets. We demonstrate the following interesting facts by two cancer research data examples.

(1) The log-log plotting cannot tell that the data do not fit Cox's regression model.

(2) Even though it seems quite obvious that the two survival functions are different, the logrank test gives insignificant results in both cases.

(3) The test based on the weighted Kaplan-Meier statistic (see Pepe and Fleming (1991)) gives significant results in both cases.

(4) Our new procedure indicates that both data do not fit Cox's regression model.

We demonstrate by another cancer research data set that, when our procedure find no evidence that Cox's model is not appropriate for that data, both the logrank test and the test based on the weighted Kaplan-Meier statistic give significant results. We further apply our new procedure to a simulation data set from Cox's regression model, and the result is not significant, as it should be.

2. Method

We shall propose a test for H_o^c: Cox's model is appropriate.

2.1. With right-censored data

Under right censoring, the observations are (M_{ij}, δ_{ij}), j = 1, ..., n_i, j = 1, 2. The GMLE of a survival function based on Sample i is

$$\hat{S}_i(t) = \prod_{M_{i(j)} \leq t} (1 - \frac{\delta_{i(j)}}{n_i - j + 1}),$$

where $M_{i(1)} \leq \cdots \leq M_{i(n_i)}$ are order statistics of M_{ij} from Sample i, and $\delta_{i(j)}$ is the δ_{ik} that assosiated with $M_{i(j)}$.

By the definition of Z and the property of the Cox model,

$$\log[-\log S_1(t|Z)] - \log[-\log S_2(t|Z)] = \beta \text{ for each } t \text{ and } Z. \tag{2.1}$$

In view of this equality, we propose to replace the log-log plot by plotting

$$U = U(t) = \log(-\log(\hat{S}_1(t))) - \log(-\log(\hat{S}_2(t))). \tag{2.2}$$

In view of (2.1), we shall inspect whether the curve is some what a horizontal straight line, that is, it is within a band. Furhtermore, in order to test H_o^c, define a set of statistics as follows. Let b_1, ..., b_{m+1} be all the distinct exact observations from the pooled sample. Compute

$$U_j = \log[-\log \hat{S}_1(b_j)] - \log[-\log \hat{S}_2(b_j)],\ j = 1, ..., m+1. \tag{2.3}$$

Thus, one expects that U_1, ..., U_{m+1} are statistically a constant. Note that if $\hat{S}_i(b_j) = 0$ or 1, $U_i = \pm\infty$. By deleting these types of b_i, without loss of generality, we can assume that U_i's are all finite. Denote $Q_i = U_{i+1} - U_i$, $i = 1$, ..., m For simplicity, we further assume that m is even (by deleting one number). Let $\mathbf{Q} = (Q_1, ..., Q_m)'$. Then Q_i has mean zero. Note that the covariance matrix Σ_1 of

$$(\hat{S}_1(t_1), ..., \hat{S}_1(t_{m_1}), \hat{S}_2(t_{m_1+1}), ..., \hat{S}_2(t_{m_1+m_2}))'$$

are known, where $t_1, ..., t_{m_1}$ are exact observations in the first group and $t_{m_1+1}, ..., t_{m_1+m_2}$ are exact observations in the second group. For convenience, denote

$$s_i = \begin{cases} \hat{S}_1(t_i) & \text{if } i = 1, ..., m_1, \\ \hat{S}_2(t_i) & \text{if } i = m_1, ..., m_1 + m_2. \end{cases}$$

Note that $S_1(b_j) = s_i$ if $t_i \leq b_j < t_{i+1}$ and $i < m_1$, or if $t_i < b_j$ and $i = m_1$; and $S_2(b_j) = s_i$ if $t_i \leq b_j < t_{i+1}$ and $i < m_1 + m_2$, or if $s_{m_1+m_2} \leq b_j$ and $i = m_1 + m_2$. By Slutsky's theorem, one can compute the covariance matrix of $\mathbf{Q}$, say, Σ. In fact, $\Sigma = H\Sigma_1 H'$, where $H = (\frac{\partial U_i}{\partial s_j})_{m \times (m_1+m_2)}$. Let $\hat{\Sigma}$ be the GMLE of Σ. Denote $\mathbf{W} = (W_1, ..., W_m)' = \hat{\Sigma}^{-1/2}\mathbf{Q}$. If n is sufficient large, say $n \geq 20$, by the asymptotic normality of $\vec{W}$, without loss of generality, one can assume

A1 W_1, ..., W_m are a random sample from a normal distribution.

Let

$$V = \frac{\sum_{i=1}^{m/2}(W_i - u_1)^2}{\sum_{i>m/2}^{m}(W_i - u_2)^2}, \text{ where } u_1 = \frac{2}{m}\sum_{i=1}^{m/2} W_i \text{ and } u_2 = \frac{2}{m}\sum_{i>m/2}^{m} W_i, \tag{2.4}$$

then we expect that V has an F distribution with $(m/2) - 1$ and $(m/2) - 1$ degrees of freedom. A test is to reject H_o^c if V is too large or too small (V should be around 1 under H_0^c).

2.2. With interval-censored data

We first discuss how to obtain a GMLE of S_i. An interval-censored observation (L_i, R_i) can be represented by an interval $I_i = \begin{cases} (L_i, R_i] & \text{if } L_i < R_i \\ [L_i, R_i] & \text{if } L_i < R_i \end{cases}$. A nonempty intersection of some I_i's, say A, is called an innermost interval if A satisfies that, for each i, $A \cap I_i$ equals $\emptyset$ or A. Turnbull (1974) shows that the GMLE based on random intervals I_i, $i = 1$, ..., n, only puts weights on innermost intervals. Let A_1, ..., A_m be all the distinct innermost intervals induced by I_1, ..., I_n. Let $\hat{s}_j$ be the weight assigned by the GMLE to A_j and $\hat{\mathbf{s}} = (\hat{s}_1, ..., \hat{s}_m)$. An GMLE of a cdf is

$$F(t) = \sum_{j=1}^{m} \hat{s}_j \mathbf{1}_{(A_j \subset (-\infty, t])}, \text{ where } \mathbf{1}_B \text{ is the indicator function of an event } B. \tag{2.4}$$

The GMLE $\hat{\mathbf{s}}$ can be obtained by the following self-consistent algorithm:

(1) Let $s_j^{(1)} = 1/m$, $j = 1, ..., m$.

(2) For $k \geq 1$, let $s_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{1}_{(A_j \subset I_i)} s_j^{(k)}}{\sum_{h=1}^{m} \mathbf{1}_{(A_h \subset I_i)} s_h^{(k)}}$, $j = 1, ..., m$.

Repeat Step 2 until convergence. The limit $\lim_{k\to\infty}(s_1^{(k)}, ..., s_m^{(k)})$ is $\hat{\mathbf{s}}$.

For more efficient algorithms for the GMLE, we refer to Wellner and Zhan (1997). By numerical methods, we can compute the GMLE's $\hat{F}$, $\hat{F}_1$ and $\hat{F}_2$, and thus can compute Λ (see (2.1)).

It is well known that for each i, given interval-censored observations (L_{ij}, R_{ij})'s, there is a unique GMLE of S_i such that it is a right-continuous step function with discontinuity points only at R_{ij}'s. Let $\hat{S}_i$ be such a GMLE of S_i, $i = 1$, 2, and let b_1, ..., b_m be all the distinct finite points at which $\hat{S}_1$ or $\hat{S}_2$ takes a jump. We propose to replace log-log plot by plotting $U = U(t)$ defined in (2.2), with new b_i's and new $\hat{S}_i$'s, and reject H_o^c if V is too large or too small, where the test statistic V is defined in (2.4) with new b_i's and new $\hat{S}_i$'s.

3. Application

In this section, we apply our new procedure to three cancer research data sets and one simulation data set. Three are right-censored data and one is interval-censored data. We demonstrate by two cancer research data examples (Example 1 and 3) the following interesting facts:

(1) The log-log plotting cannot tell that the data do not fit Cox's regression model.

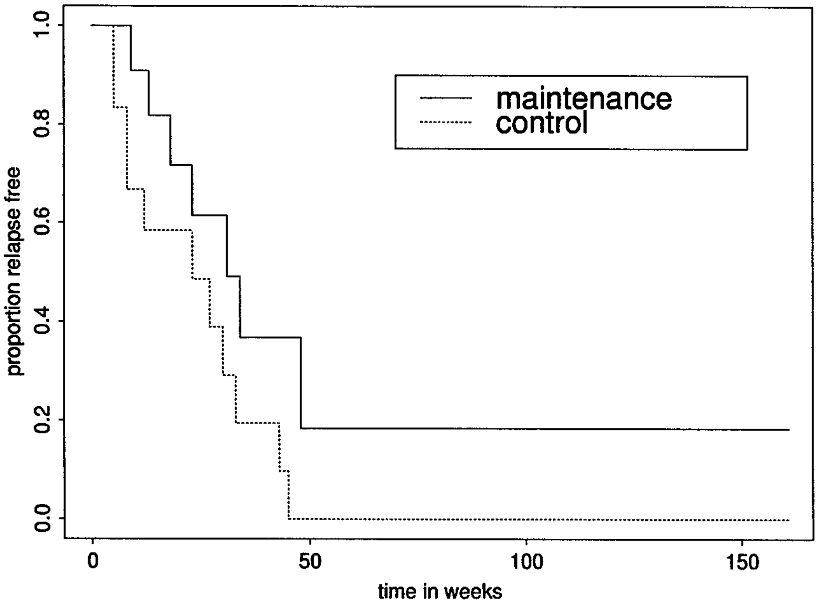
(2) Even though it seems quite obvious from survival plots that the two survival functions are different and The test based on the weighted Kaplan-Meier statistic (see Pepe and Fleming (1991)) suggests that the difference is significant in both cases, the logrank test suggests there is no difference in both cases.

(3) Our new procedure indicates that both data do not fit Cox's regression model.

We demonstrate by Example 2 that when our procedure find no evidence that Cox's model is not appropriate for that data, both the logrank test and the test based on the weighted Kaplan-Meier statistic give significant results. Finally, we generated two samples of data from Cox's regression model, our new test reveals that there is no evidence that the data are not from Cox's model, as we expect.

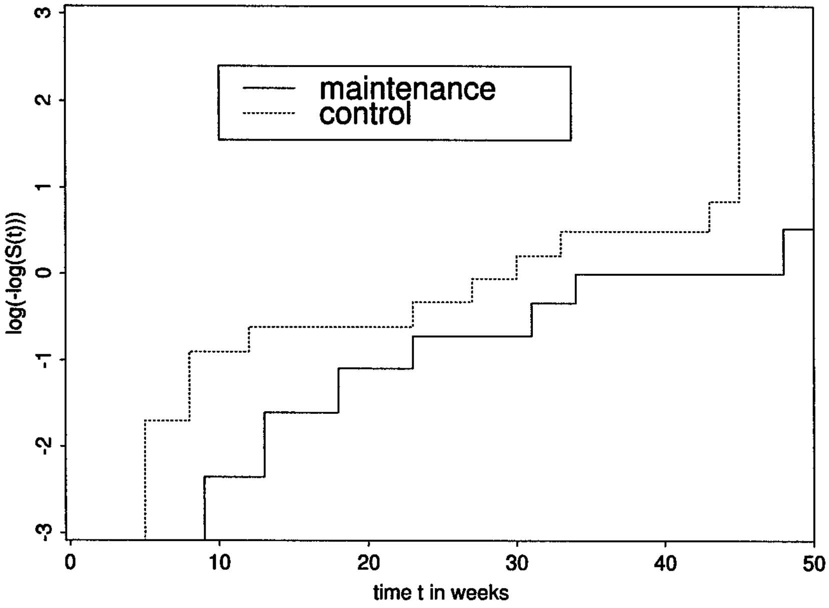
Example 1. (the AML Maintenance Study) (see Miller (1980)).

Fig. 1. Survival PLot for Leukemia Data



A clinical trial to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukemia (AML) was conducted. There are 11 right-censored data in the maintained group: 9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+. There 12 data in the control group: 5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45. The GMLE's of the survival functions of these two groups are plotted in Figure 1. One may want to test whether the two distribution are the same. In Figure 2, we plot $\log(-\log(\hat{S}_i(t)))$. It seems that the two curves are parallel. The logrank test gives a P-value 0.0655. That is, the two distribution functions are not significant different.

Fig. 2. log-log Survival PLot for Leukemia Data

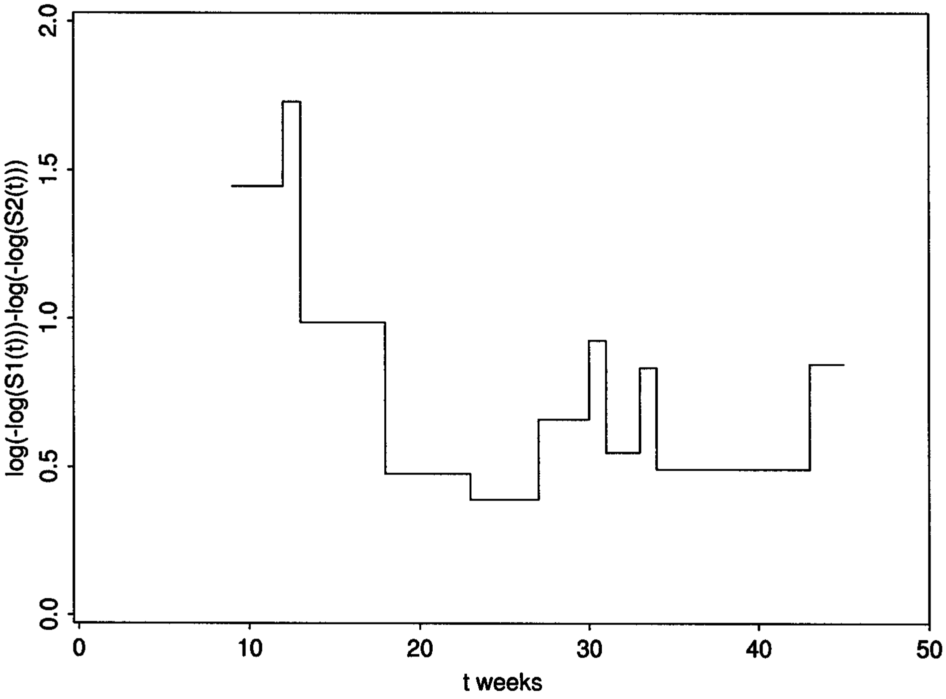


Pepe and Fleming (1991) propose a test based on the statistic $\int_0^T [\hat{S}_1(t) - \hat{S}_2(t)]dt$, where T is the longest follow-up time. For this data set, it gives a P-value < 0.01. Thus the test suggests that the two distributions are significantly different, which is consistent with Figure 1.

We plot the curve $\log(-\log(\hat{S}_1(t))) - \log(-\log(\hat{S}_2(t)))$ in Figure 3. The curve in Figure 3 does not appear to be a band. Here $S_1(t)$ is the survival function of the control group and $S_2(t)$ is the survival function of the maintenances group. We further compute statistics U_i's and V. For the current data set, U_i's equal ∞, ∞, 1.4467262, 1.7316298, 0.9870430, 0.4784492, 0.3912877, 0.6607606, 0.9265938, 0.5493081, 0.8340220, 0.4937044, 0.8466185, ∞, ∞.

Note that for this data set, $m = 15$, $U_i = \infty$ at $i = 1$ or 2 and $U_i = 0$ at $i = 14$ or 15. Thus we delete these four points. By further deleting the middle point (so that $m = 10$), we found that V is extremely large ($= 9.1$), with a P-value < 0.025. If we delete the 13rd point, $V = 10.1$, the P-value is also < 0.025. Thus we concluded that it is unlikely that the data fit the Cox model. Thus it is not a surprise that the logrank test does not reject H_0.

Fig. 3. Diagnostic PLot for Leukemia Data



Example 2. Survival data of 30 patients with AML are given in Lee (1992,p.257). Among them 17 patients are old ($\geq$ 50 years) with survival times: 6, 7, 8, 9, 15, 18, 19+, 23, 28+, 28+, 31, 39+, 45+. The survival times of the younger group are 2, 3, 3, 3, 4, 4, 8, 8, 9, 10, 12, 13, 13, 14, 18, 26+, 35+. The two GMLE's of the survival functions are plotted in Figure 4.

In Figure 5, $\log(-\log(\hat{S}i))$ are plotted. Lee claims that Figure 5 suggests that the data fit the Cox regression model. We use our approach to justify that claim. Here $S_1(t)$ is the survival function of the older group and $S_2(t)$ is the survival function of the younger group.

Fig. 4. Survival PLot for AML Data

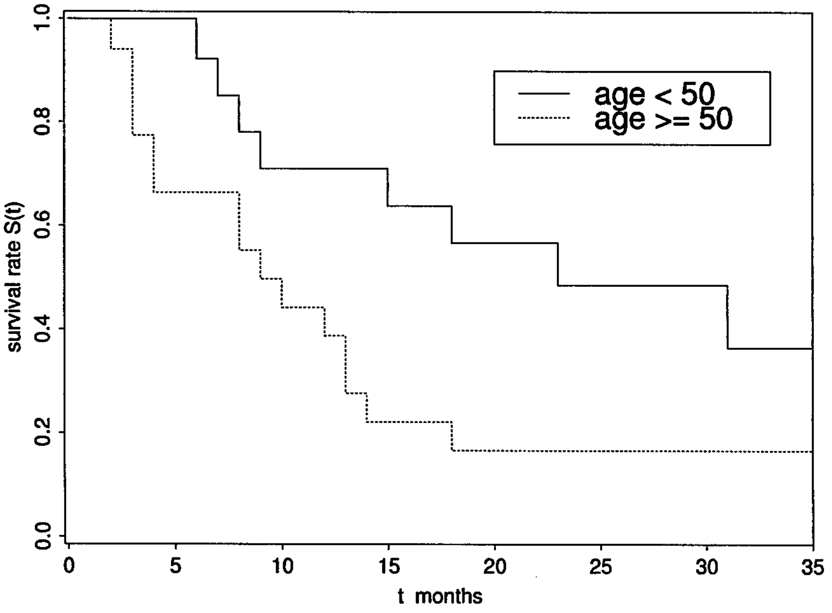
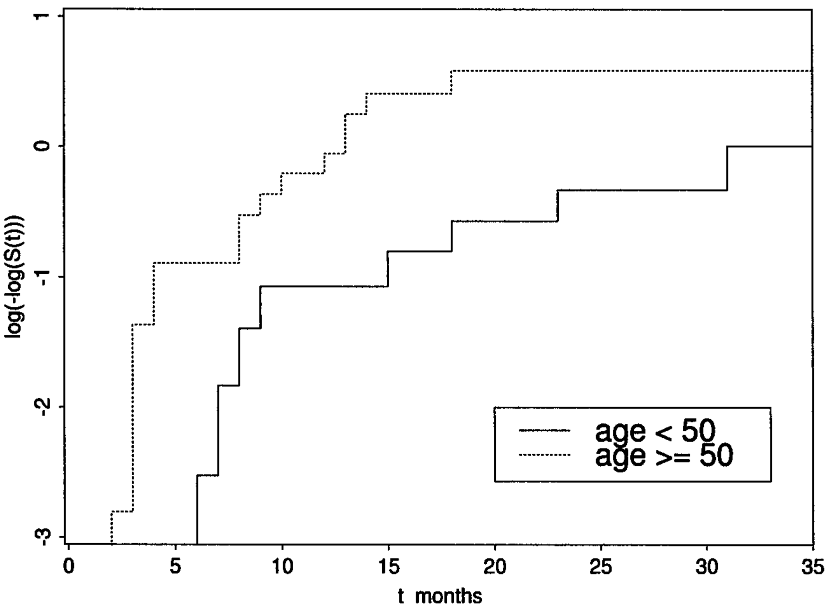


Fig. 5. log-log Survival PLot for AML Data

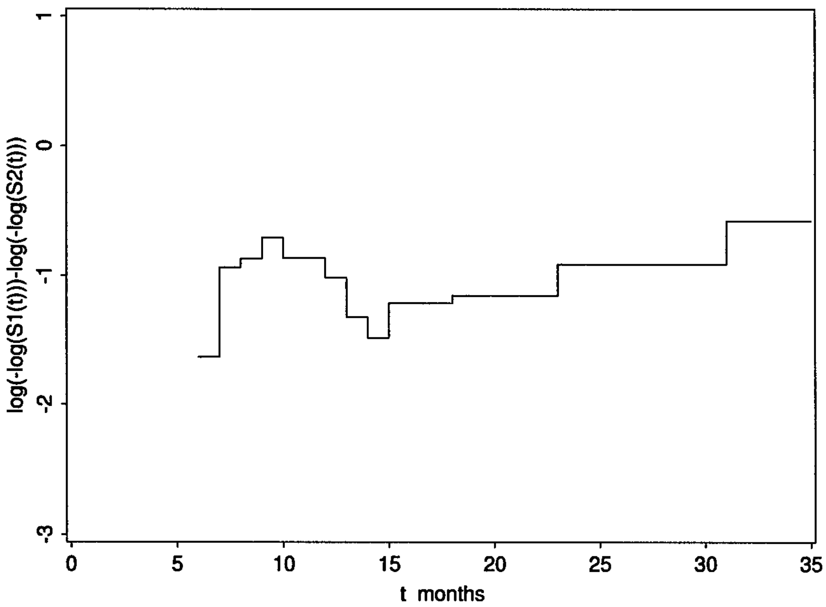


We plot the curve $\log(-\log(\hat{S}_1(t))) - \log(-\log(\hat{S}_2(t)))$ in Figure 6. The curve in Figure 6 does appear to be a band.

$V = 1.00419$ with 5 and 5 degrees of freedom, with a P-value > 0.1. Thus there is no evidence that the Cox Model is not appropriate.

The logrank test has a P-value 0.0249. This is an example that if the data fit the Cox model then the logrank test is very powerful.

Fig. 6. Diagnostic PLot for AML Data



Example 3. We generate two samples of right-censored data from exponential distributions, with density functions $f(x:\theta) = \theta e^{-x/\theta}$, $x > 0$, $\theta = 1$, 2. The first sample has 20 data:

0.0115, 0.1240, 0.1660, 0.2830+, 0.3286, 0.3603, 0.4047, 0.4586, 0.4599, 0.4762,

0.8303, 0.8379+, 0.8647, 1.0008, 1.0872+, 1.1345, 1.1740, 1.2917+, 1.6129, 2.7834+,

The second sample also has 20 data:

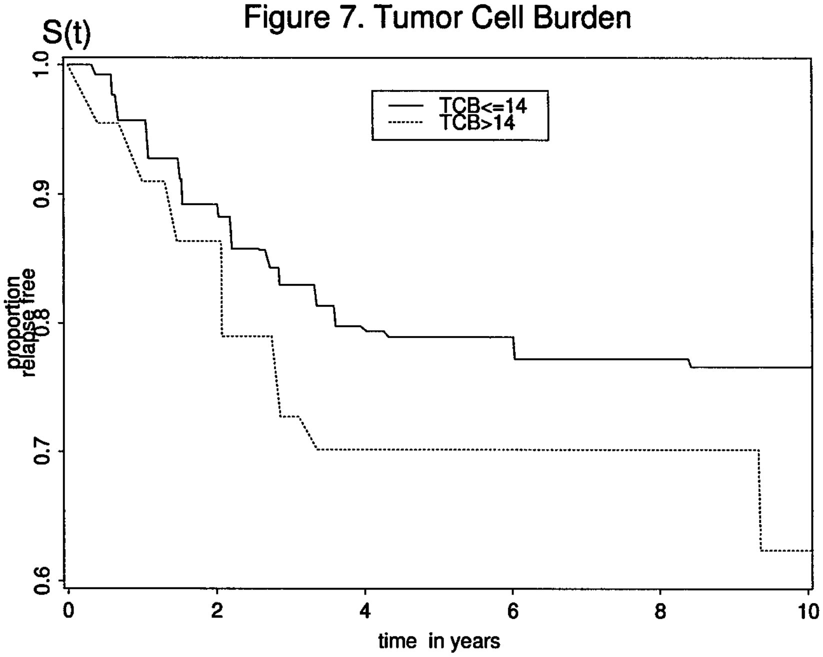
0.0089, 0.0622, 0.0722, 0.1307, 0.1345, 0.1751, 0.2390, 0.2707, 0.4198, 0.4489,

0.4564, 0.5426, 0.6066, 0.7225+, 0.8942, 0.9409, 0.9886, 1.4954, 1.9492, 2.4567.

The test statistic V has a P-value > 0.05. Thus we cannot reject the null hypothesis H_o^c, which is a correct decision in this simulation study.

Example 4. We applied the new procedure to a standard breast cancer relapse follow-up study based on data from 375 women with stages I - III unilateral invasive breast cancer surgically treated at Memorial Sloan-Kettering Cancer Center between 1985 and 1990. The median follow-up duration was 46 months. Relapse time was given by the time interval between surgery and the initial relapse. A relapse that took place between two successive follow-up visits was regarded as interval censored. If a patient did not relapse toward the end of the study, then her relapse time was right censored. Of the 375 observations, 300 were right censored (no relapse), 21 were left censored and 54 were strictly interval censored. Bone marrow micrometastasis (BMM) was determined for each woman at the time of surgery. An important question is whether remission duration is related to the extent of initial tumor cell burden (TCB) defined as number of BMM cells detected. We grouped the patient according to whether the patient's number of BMM cells is ≤ 14 or > 14.

Figure 7. Tumor Cell Burden



The GMLE's of the survival functions of these two groups are plotted in Figure 7. It seems from the figure that the two survival functions are different. We also given the log-log plot in Figure 8. It is hard to say from the figure that the two curves are not parallel. Thus we let Z be the indicator function of the event that the observation is from Population 2, *i.e.*, the number of BMM cells is > 14, and assume that the data fit Cox's regression model. It turns out that the semi-parametric MLE of β is 0.328 with a standard deviation 0.293. However, the P-value is 0.131. That is, β is not significantly different from 0. The Cox model approach says that the BMM has no effect on the survival rate. The conclusion is not consistent with Figure 1.

Figure 8. Log-Log Plot For Tumor Cell Burden

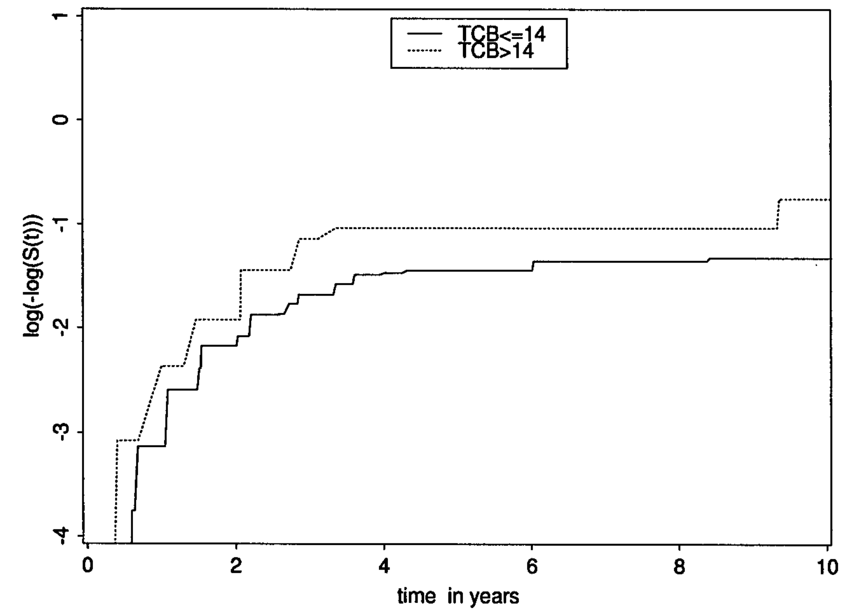
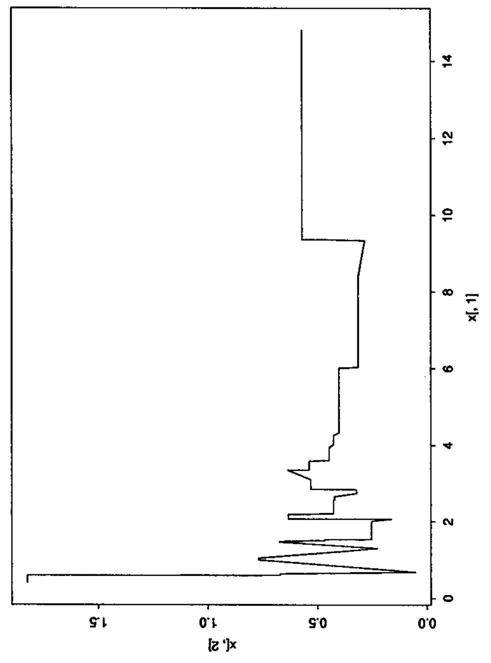


Figure 8. Log-Log Difference Plot For TCB



We compute the test based on the statistic $\int W(t)(\hat{F}_1(t) - \hat{F}_2(t))dt$, where $W(t) = \mathbf{1}_{(0,T)}$ and T is a constant. The P-value is 0.029. In our calculation, we chose $T = 3500$, the longest follow-up time of a patient. The result indicates that the effect of BMM is significant. That is, H_0 is not true.

Applying the new approach to the data, we found that V is extremely large, with a P-value < 0.01. Thus we concluded that it is unlikely that the data fit the Cox model. Thus it is not surprised why the Cox regression analysis performs poorly.

We further check whether the grouping in BMM and BMM- would fit the PH model. The P-value is 0.013 and the diagonostic plots are give in Figure 9.

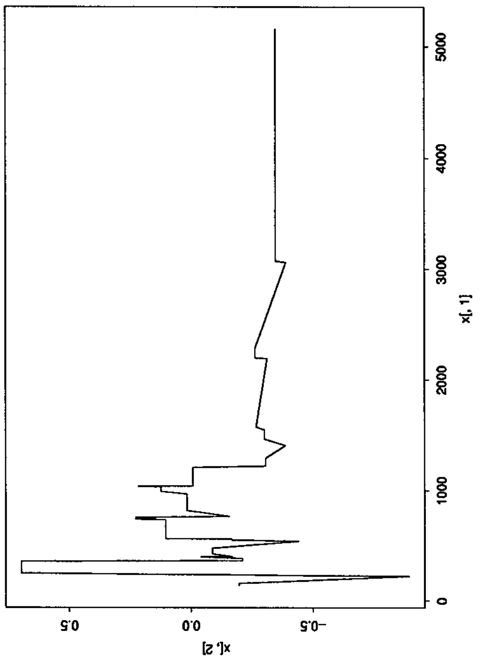


Figure 9. Log–Log Difference Plot For BMM

References

- * Cox, D.R. and Oakes, D. (1984). Analysis of survival data. *Chapman & Hall.* London.
- * Lee, E.T. (1992). Statistical methods for survival data analysis. *Wiley.* N.Y.
- * Miller Jr., R. G. (1981). Survival analysis. *Wiley.* p. 146.
- * Pepe, M. S. and Fleming T. R. (1991). Weighted Kaplan-Mieir Statistics: Large sample and optimality considerations. *J.R.S.S. B*, 53, 2. 341-352.
- * Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *JASA* 69 169-173.
- * Wellner, J.A. and Zhan, Y. (1997). A hybrid algorithm for computation of the NPMLE from censored data. *JASA* **92** 945-959.

Reference [12]

"Communication in Statistics-Simulation and Computation" 32 p.147-164 (2003)

SEMI-PARAMETRIC MLE IN SIMPLE LINEAR REGRESSION ANALYSIS WITH INTERVAL-CENSORED DATA

Qiqing Yu * and George Y. C. Wong *

Mathematics Department, SUNY at Binghamton, NY 13902, USA
E-mail: qyu@math.binghamton.edu
and
Strang Cancer Prevention Center, 428 E 72nd Street, NY 10021, USA

Running title: SMLE in linear regression.

Current version 3/12/2002
AMS 1991 subject classification: Primary 62 J05; Secondary 62 G05.
Key words: bootstrap, generalized likelihood, iterative algorithm, consistency.

ABSTRACT

Consider the model $Y = \beta X + \epsilon$ with interval-censored data, where ϵ has an unknown cdf F_o. The semi-parametric MLE (SMLE) of β is well defined, but cannot be obtained by algorithms for M-estimators, or by the Newton-Raphson method or the Monte-Carlo method. Thus it has not been studied in the literature even in the case of complete data. We propose a feasible algorithm to obtain all solutions of the SMLE. Simulation suggests that the SMLE is consistent and the bootstrap estimator of the variance of the SMLE matches the sample variance. We compare the SMLE to the Buckley-James estimator (BJE) in four data sets with sample sizes up to 374. The results show that the SMLE is more robust and more reliable than the BJE.

1. INTRODUCTION

Regression analysis is one of the most widely used statistical techniques. Its applications occur in almost every field, including economics, engineering, the physical sciences, management, life and biological sciences and the social sciences. We consider the simple linear regression problem with interval-censored data. In particular, we assume that

A1. $Y = \beta X + \epsilon$, where Y is a random variable, X a covariate taking at least two values, β an (unknown) regression coefficient and ϵ a random variable with an unknown c.d.f. F_o.

Since no further assumption is made on (β, F_o), this is a semi-parametric problem.

There are several estimators for the simple linear regression model with complete data. They are the least squares estimator (LSE), Theil's estimator (Theil (1)), various M-estimators (Huber (2)), adaptive estimators (Bickel (3)), and Bayes estimators.

Sometimes Y may be right censored. There have been extensive studies on linear regression models with right-censored data, see, *e.g.*, Buckley and James (4), Miller (5) and Ritov (6), among others.

* Partially supported by DAMD17-99-I-9390 and DAMD17-00-1-0448.

Interval censoring is another type of censoring. For example, in medical follow-up studies, a patient may be inspected (or interviewed) total of K times ($K \geq 1$), at $C_{K,1} < \cdots < C_{K,K}$. Time to event of interest, Y, is often not observable but is only known to lie within two consecutive inspection times $C_{K,j}$ and $C_{K,j+1}$, where $C_{K,0} = -\infty$, $C_{K,K+1} = \infty$, and $1 \leq j \leq K$. In such a case we are only able to observe (L, R, X), where $-\infty \leq L < Y \leq R \leq +\infty$. We say such an observation is *interval censored.* Hereafter, we assume that

A2. the observable random vector is (L, R, X), where (L, R) is an extended random vector (*e.g.*, $R = \infty$ under right censoring) such that either $L < Y \leq R$ or $L = Y = R$.

Odell, *et al.* ((7)) considered the parametric maximum likelihood estimation for interval -censored data based on a Weibull distribution. Self and Grossman (8) considered the linear regression problem with interval-censored data for a given distribution F_o with location-scale parameter and proposed a marginal likelihood approach.

Several authors have proposed estimators for β under linear regression models with interval-censored data, assuming F_o is unknown. Rabinowitz, *et al.* (9) proposed a class of score statistics to estimate β. Their approach parallels the construction of the Buckley & James estimator (BJE) for right-censored data. Li and Pu (10) considered generalizations of Miller's estimator and the BJE for interval-censored data that contain exact observations. Zhang and Li (11) and Li and Zhang (12), among others, studied M-estimators with doubly-censored data and Case 1 interval-censored data. These approaches can be viewed as a modification of the Semi-parametric MLE (SMLE). But they are not an SMLE (see Example 1.1 below).

Under the semi-parametric set-up, it is well known that the LSE is not efficient unless F_o is a normal distribution. The BJE is an extension of the LSE under censoring (see Li and Zhang (12)), thus the BJE is not efficient. Consequently, since the BJE is also an M-estimator (see Zhang and Li ((11), p. 2723)), an M-estimator may not be efficient. Li and Zhang proposed efficient M-estimators for Case 1 interval-censored data. However, it is not clear how to obtain an efficient estimator for arbitrary interval-censored data.

The SMLE of (β, F_o) has long been ignored and there is no algorithm in the literature for obtaining the SMLE with interval-censored data. Recently, Li and Zhang (12) mentioned without a proof that the SMLE (they called the profile MLE) should be consistent with Case 1 interval-censored data. However, there is no hint on how to compute it. Unlike the MLE in other cases, the SMLE cannot be computed by standard numerical methods, *e.g.*, the Monte Carlo method, the Newton-Raphson method, the M-estimation methods and the finite algorithms discussed in Osborne (13). For the sake of simplicity, we illustrate with 4 artificial complete observations as follows.

Example 1.1. Suppose the observations (X_i, Y_i)'s are: $(-1, 0)$, $(0, 1)$, $((1, 2)$ and $(4, 0)$. The generalized likelihood function (Kiefer and Wolfowitz (14)) is

$$Ł(F, b) = \prod_{i=1}^{4} f(Y_i - bX_i), \text{ where } f(t) = F(t) - F(t-) \text{ and } F \text{ is a cdf.} \tag{1.1}$$

Then, given b, Ł is maximized by the empirical cdf $\hat{F}_b$. where $\hat{F}_b(t) = \frac{1}{4}\sum_{i=1}^{4} \mathbf{1}_{(Y_i - bX_i \leq t)}$ and $\mathbf{1}_A$ is the indicator function of the event A. That is,

$$Ł(F, b) \leq Ł(\hat{F}_b, b) \text{ and } Ł(\hat{F}_b, b) = \begin{cases} (\frac{3}{4})^3 \frac{1}{4} & \text{if } b = 1, \\ (\frac{1}{2})^2 (\frac{1}{4})^2 & \text{if } b = 0, \ -1/4 \text{ or } -2/3, \\ (\frac{1}{4})^4 & \text{otherwise.} \end{cases} \tag{1.2}$$

Thus the SMLE of β is 1.

A key requirement in the Newton-Raphson method and the Monte Carlo method (see, *e.g.*, Ingber (15)), as well as in the finite algorithms discussed in Osborne (13), is that Ł is continuous at a neighborhood of the maximum point or Ł is convex. It follows from (1.2) that $l(b) = Ł(\hat{F}_b, b) = \frac{1}{128}$ a.e., which is not the maximum of $Ł(F, b)$. That is, $l(\cdot)$ is neither continuous at the SMLE nor is convex in b. Consequently, the two standard methods, as well as the finite algorithms discussed in Osborne, do not help to find the SMLE.

The M-estimate considered by Zhang and Li (12) is a solution to $\frac{\partial \ln Ł(F,b)}{\partial b} = 0$, where F is properly defined. Since $\frac{d \ln l}{b} = 0$ a.e. by (1.2), the SMLE is somewhat an M-estimator. But this M-estimation approach is non-informative, as every value of b is an M-estimate. Theil's estimator is the median of the collection of slopes of the line segments connecting (X_i, Y_i) and (X_j, Y_j), where $1 \leq i < j \leq n$ $(= 4)$. Thus, Theil's estimator is not an SMLE. It is easy to verify that the LSE is not an SMLE neither. □

Remark 1.1. It is worth mentioning that in Example 1.1, one may modify the likelihood function in (1.1) as follows:

$$\mathrm{L} = \prod_{i=1}^{n} f(Y_i - bX_i), \ f = F' \text{ and } F \in \Theta_0, \tag{1.3}$$

where Θ_0 is a subset of certain absolutely continuous cdf's. Then the values of (F, b) that maximizes L over all b and over all $F \in \Theta_0$ can be called a modified SMLE or repaired SMLE. But it has not been called the SMLE in the literature. This modification is the motivation of the M-estimation approach and the approach based on score functions.

More recently, Yu and Wong (16) studied the SMLE with right-censored data. Simulation suggests that the SMLE $\hat{\beta}$ is consistent and $\lim_{n\to\infty} nVar(\hat{\beta}) = 0$ if F_o is discontinuous. The property was proved under a discrete assumption that (X, Y) takes on finitely many values. In contrast, all the existing estimators do not have the second property. Under continuous assumptions, the SMLE with interval-censored data may attain the efficient lower bound (see Cosslett (17)).

In this paper, we shall study how to derive an SMLE with interval-censored data. In Section 2, we define the SMLE. In Section 3, we propose feasible algorithms to obtain all possible SMLE's. In Section 4, we present simulation results. The simulation results indicate that the SMLE is computationally feasible, is consistent and its standard error can be estimated by the bootstrap method. In Section 5, we apply our procedure to four data sets and compare to the LSE or the BJE. Several comments are made in Section 6.

2. THE SMLE

Let $(Y_i, X_i, \epsilon_i, L_i, R_i)$, $i = 1, \ldots, n$, be i.i.d. copies of (Y, X, ϵ, L, R). Denote random intervals

$$I_i = I_i(b) = \begin{cases} (L_i - bX_i, R_i - bX_i] & \text{if } L_i < R_i, \\ [L_i - bX_i, R_i - bX_i] & \text{if } L_i = R_i. \end{cases}$$

Note that I_i is a singleton if $L_i = R_i$. Denote μ_F the measure induced by $F \in \mathcal{F}$, where $\mathcal{F}$ is the class of all distribution functions. In other words,

$$\mu_F(I) = \begin{cases} F(v) - F(u) & \text{if } I = (u, v], \\ F(v) - F(u-) & \text{if } I = [u, v]. \end{cases}$$

By assumptions A1 and A2, $I_i(\beta)$ are i.i.d. random intervals and $\epsilon_i \in I_i(\beta)$. Thus the generalized likelihood function defined by Kiefer and Wolfowitz (14) is

$$\mathrm{L}(F, b) = \prod_{i=1}^{n} \mu_F(I_i(b)), \ F \in \mathcal{F}, \ b \in \mathcal{R} \ \ (= (-\infty, \infty)). \tag{2.1}$$

A semi-parametric MLE of (F_o, β) maximizes L. Given b, the GMLE of F_o based on observations $I_i(b)$'s maximizes $\mathrm{L}(\cdot, b)$ over $F \in \mathcal{F}$. Thus, in order to find the SMLE of (F_o, β), it suffices to maximize

$$l(b) = \mathrm{L}(\hat{F}_b, b), \ \ b \in \mathcal{R}. \tag{2.2}$$

The SMLE of β may not be unique (see Example 3.1). Denote the set of all solutions of the SMLE of β by $\mathcal{B}$. Then each $\hat{F}_b$, $b \in \mathcal{B}$, is an SMLE of F_o.

Under assumption A1, $\alpha = E(\epsilon)$ may not exist. If it does exist, α can be estimated by $\hat{\alpha} = \int t\hat{F}_b(t)$, where $b \in \mathcal{B}$. However, even if $b = \beta$, $\hat{\alpha}$ is not consistent unless Y is observable (with positive probability) everywhere on its range, or P(Y is not censored$|Y = t) > 0$ for all possible t. Thus, in general, α is not identifiable under censoring (see, *e.g.*, Buckley and James (4)). We shall ignore $\hat{\alpha}$ in our study.

3. METHODS

In this section, we shall first introduce an algorithm which guarantees to obtain all solutions of the SMLE and then introduce another algorithm which is faster, but offer no proof.

In view of (2.2), in order to find the SMLE, it suffices to compare $l(b)$, $b \in \mathcal{R}$. Let $u_i(b)$ and $v_i(b)$ be the endpoints of the interval $I_i(b)$, *i.e.*, $u_i(b) = L_i - bX_i$ and $v_i(b) = R_i - bX_i$. Let $T_{2i-1}(b) = u_i(b)$ and $T_{2i}(b) = v_i(b)$, $i = 1, ..., n$. Then we can determine the ranks of the 2n extended random variables. Given b, it is well known that $\mathrm{L}(\hat{F}_b, b)$ only depends on the ranks of the 2n $T_i(b)$'s (see Turnbull (18)). That is,

$$l(b_1) = l(b_2) \text{ if the rank of } T_i(b_1) \text{ is the same as the rank of } T_i(b_2) \text{ for each } i. \tag{3.1}$$

Note that the ranks of these $T_i(b)$'s will change only at the solutions of the equations $T_i(b) = T_j(b)$, $i \neq j$. The latter equations yield equations of forms

$$L_i - bX_i = L_j - bX_j,\ L_i - bX_i = R_j - bX_j \text{ or } R_i - bX_i = R_j - bX_j, \tag{3.2}$$

where L_i, R_i, L_j and R_j are finite, and $X_i \neq X_j$. Since there are at most $4n^2$ equations of forms in (3.2), there are at most $4n^2$ distinct solutions to these equations, denoted by $b_1 < \cdots < b_m$. Let $b_0 = -\infty$ and $b_{m+1} = \infty$. By construction, if $b \in (b_k, b_{k+1})$, then $T_i(b)$'s will not change their ranks. Consequently, it follows from (3.1) that

$$\text{for each } k,\ l(b)\ (= \mathrm{L}(\hat{F}_b, b)) \text{ is constant on the open interval } (b_k, b_{k+1}). \tag{3.3}$$

There are $m+1$ disjoint open intervals of form (b_k, b_{k+1}) and m disjoint closed intervals of form $[b_k, b_k]$. As a consequence, there are at most $2m+1$ distinct values of $l(b)$, which can be represented by $l(b_j^o)$, $j = 1, ..., 2m+1$, where $b_1^o = b_1 - 1$, $b_{2m+1}^o = b_m + 1$, $b_{2i}^o = b_i$, $i = 1, ..., m$, and $b_{2i+1}^o = (b_i + b_{i+1})/2$, $i = 1, ..., m-1$. Denote

$$\mathcal{A}_1 = \{b_1^o, ..., b_{2m+1}^o\}. \tag{3.4}$$

In order to find an SMLE of β, it suffices to compare $l(b)$ for $b \in \mathcal{A}_1$. Let $\mathcal{B}^o$ be the set of all points in $\mathcal{A}_1$ that maximize $l(b)$, $b \in \mathcal{A}_1$. The $\mathcal{B}^o \subset \mathcal{B}$. Moreover, if $b_{2j+1}^o \in \mathcal{B}^o$, then $(b_{2j}^o, b_{2j+2}^o) \subset \mathcal{B}$ by (3.3). To summarize, we have the following algorithm:

Algorithm 3.1 (for obtaining all solutions of the SMLE of β):

(1) Derive the set $\mathcal{A}_1$ (see (3.4)).

(2) Compute $\hat{F}_b$ and $l(b)$, where $b \in \mathcal{A}_1$. For computing $\hat{F}_b$, we refer to Turnbull (18) or Wellner and Zhan (19). Then

$$b \in \mathcal{B} \text{ iff either (1) } b \in \mathcal{B}^o \text{ or (2) } b \in (b_{2j}^o, b_{2j+2}^o) \text{ for a } b_{2j+1}^o \in \mathcal{B}^o.$$

Each $(\hat{\beta}, \hat{F}_{\hat{\beta}})$, $\hat{\beta} \in \mathcal{B}$, is an SMLE of (β, F_o).

If the SMLE is not unique, one can make the following choices: (1) choosing an SMLE that is closest to the median of $\mathcal{B}$; (2) choosing an SMLE that is closest to the median of $\mathcal{B}^o$; (3) choosing an SMLE that is closest to the BJE. The second choice is the easiest one to implement.

It is often time-consuming to compute the GMLE $\hat{F}_b$. Thus it is desirable to reduce the distinct values of b involved in Step 2 of Algorithm 3.1. The following algorithm reduces the number of b involved by a factor of 4.

Algorithm 3.2.

1. Find the solution b to the equation of form $L_i - bX_i = L_j - bX_j$ or $R_i - bX_i = R_j - bX_j$, where L_i, L_j, R_i and R_j are finite, and $X_i \neq X_j$. Let $\mathcal{A}_2$ be the set of all the distinct elements of these solutions.
2. Compute $\hat{F}_b$ and $l(b)$, $b \in \mathcal{A}_2$. Then identify those points in $\mathcal{A}_2$ which maximize $l(b)$, $b \in \mathcal{A}_2$ and denote $\mathcal{B}^*$ the collection of these points.
3. Choose $b \in \mathcal{B}^*$ that is closest to the median of $\mathcal{B}^*$. Treat it as the SMLE of β.

This algorithm is faster and our simulation results suggest that if n is large then it yields an SMLE. To accelerate the algorithm, we may further make the following modifications:

(i) Randomly select a subset J of the set $\{(i,j) : 1 \leq i < j \leq n\}$. Let $\mathcal{A}_2^*$ be the collection of the solutions b to the equation of form $L_i - bX_i = L_j - bX_j$ or $R_i - bX_i = R_j - bX_j$, where $(i,j) \in J$.

(ii) Modify $\mathcal{A}_2$ as $\mathcal{A}_3 = \mathcal{A}_2^* \cap [a,b]$, where $[a,b]$ is an interval to which we suspect that β belongs. For example, we may let $[a,b] = [\tilde{\beta} - 3\tilde{\sigma}_{\tilde{\beta}}, \tilde{\beta} + 3\tilde{\sigma}_{\tilde{\beta}}]$, where $\tilde{\beta}$ is the BJE and $\tilde{\sigma}_{\tilde{\beta}}$ is the standard error (SE) of $\tilde{\beta}$.

Note that Step (i) reduces the amount of (i, j) in calculation, and Step (ii) further reduces the cost in computing the GMLE of F for selected $b \in \mathcal{A}_3$.

The following example illustrates the difference between the two algorithms.

Example 3.1. Suppose that there are 3 observations (L_i, R_i, X_i)'s. They are $(1, 4, 1)$, $(3, 6, -1)$, $(1, 2, 1)$. We first consider Algorithm 3.1. Step 1 results in $m = 5$, $(b_1, b_2, b_3, b_4, b_5) = (-2.5, -2, -1, -0.5, 0.5)$ and

$\mathcal{A}_1 = \{b_1^o, \ldots, b_{11}^o\} = \{-3.5, -2.5, -2.25, -2, -1.75, -1, -0.75, -0.5, 0, 0.5, 1.5\}$.

In Step 2, we need to compute $\hat{F}_b$ and $l(b)$, $b \in \mathcal{A}_1$. For simplicity, we only demonstrate for b_1^o and b_3^o.

(1) $b_1^o = -3.5$. Then the I_i's are $(4.5, 7.5]$, $(-0.5, 2.5]$, $(4.5, 5.5]$. $\hat{F}_{-3.5}(t) = \frac{1}{3}\mathbf{1}_{(t \geq 1)} + \frac{2}{3}\mathbf{1}_{(t \geq 5)}$. $l(b_1^o) = \frac{1}{3}(\frac{2}{3})^2$.

(2) $b_3^o = -2.25$. Then the I_i's are $(2.25, 6.25]$, $(0.75, 3.75]$ and $(2.25, 4.25]$. $\hat{F}_{-2.25}(t) = \mathbf{1}_{(t \geq 3.5)}$ and $l(b_3^o) = 1$.

To summarize, Step 2 gives the following results:

$$\begin{array}{lccccccccccc}
i: & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
b_i^o: & -3.5 & -2.5 & -2.25 & -2 & -1.75 & -1 & -0.75 & -0.5 & 0 & 0.5 & 1.5 \\
l(b_i^o): & \frac{4}{27} & \frac{4}{27} & 1 & 1 & 1 & 1 & 1 & 0.25 & 0.25 & 0.25 & 0.25
\end{array}$$

Thus $\mathcal{B} = (-2.5, -0.5)$ and $\mathcal{B}^o = \{-2.25, -2, -1.75, -1, -0.75\}$. In this example there are infinitely many SMLE's. We choose the median of $\mathcal{B}^o$, that is, $\hat{\beta} = -1.75$.

If we use Algorithm 3.2, Step 1 results in $\mathcal{A}_2 = \{-2, -1\}$ and Step 2 results in an estimate $\hat{\beta} = -1$. Both algorithms lead to an SMLE and the second algorithm is obviously faster.

Remark 3.1. Variance estimation is important for making inferences. Since we do not make any specific assumption on F_o, there are two possibilities for the variance of $\hat{\beta}$.

1. Under interval censoring without exact observations, if the regularity conditions stated for the Cramer-Rao lower bound hold and F_o is differentiable, then a consistent estimator of the efficient lower bound for the regression problem is

$$\hat{\sigma}_{\hat{\beta}}^2 = 1 / \sum_{i=1}^{n} \left(\frac{f(R_i - \beta X_i) - f(L_i - \beta X_i)}{\mu_F(I_i(\beta))} (X_i - \overline{X}) \right)^2 \Big|_{\beta = \hat{\beta}, f(t) = \frac{F(t+h_n) - F(t-h_n)}{2h_n}, F = \hat{F}_{\hat{\beta}}}, \qquad (3.5)$$

where $h_n \approx 0$, *e.g.*, $h_n = n^{-1/5}$.

2. If F_o is differentiable but $\frac{\partial E(\ln l(\beta))}{\partial \beta} \neq E(\frac{\partial \ln l(\beta)}{\partial \beta})$, or if F_o is not continuous, then the regularity conditions stated for the Cramer-Rao lower bound do not hold. In such situations, $\hat{\sigma}_{\hat{\beta}}$ is not a good estimator.

In view of the above discussion, we suggest to use the bootstrap method (see, *e.g.*, Davison and Hinkley (20)) to estimate the variance of $\hat{\beta}$. It seems to us from simulation that the bootstrap estimator is consistent.

4. SIMULATION RESULTS

Hereafter, we denote $\hat{\beta}$ the SMLE. The simulation results suggest that $\hat{\beta}$ is consistent and the bootstrap estimates of the variances of $\hat{\beta}$ match the sample variances. The simulation results also indicate that the SMLE is a feasible procedure computationally.

In our simulations, we make use of a Case 1 or Case 2 interval censorship model. Under the Case 2 interval censorship model, the observable random vector satisfies

$$(L, R) = \begin{cases} (-\infty, U) & \text{if } Y \leq U \\ (U, U+V) & \text{if } U < Y \leq U + V \\ (U+V, \infty) & \text{if } Y > U + V, \end{cases}$$

$V > 0$, and Y and (U, V) are independent. Under the Case 1 interval censorship model, the observable random vector satisfies

$$(L, R) = \begin{cases} (-\infty, U) & \text{if } Y \leq U \\ (U, \infty) & \text{if } Y > U, \end{cases}$$

and Y and U are independent.

In our simulation, we further assume that the underlying distribution of Y and the censoring vector satisfy the following conditions: (1) U is a continuous nonnegative random variables; (2) V is a nonnegative continuous random variable; (3) X, ϵ, U and V are independent.

Note that in our assumptions, ϵ and X may be continuous or discontinuous. We present simulation results under three different sets of distributions. Under each set of assumptions, we compute $\hat{\beta}$ based on sample sizes 30 and 100, respectively, with 500 simulations in each case. The program was written in C language and the simulation was carried out on a Pentium III PC.

Case 1. (Case 2 interval censoring). Assume that ϵ has the uniform distribution on the interval $(0, 2)$; U, V and X have exponential distributions; and $\beta = 1$.

Case 2. (Case 2 interval censoring). Assume that ϵ takes values 0.8 and 4 w.p. 0.8 and 0.2, respectively; X is a discrete random variable which takes values i w.p. $i/15$, $i = 1, 2, 3, 4, 5$; $\beta = 1$; and U and V have uniform distributions on the intervals $(0, 4)$ and $(0.5, 5)$, respectively.

Case 3. (Case 1 interval censoring). Assume that ϵ has a uniform distribution on the union of intervals $(0, 0.5) \cup (50.5, 51)$; the inspection time U has a uniform distribution in the interval (0,3); X takes values 0.5 and 1 w.p. 0.5 and 0.5, respectively; and $\beta = 1$.

Table 1. Simulation Results on estimating β with interval-censored data.

cases		$n = 30$	$n = 100$	β
Case 1.	$\hat{\beta}_n$ average (SE)	1.020 (0.710)	1.060 (0.401)	1
continuous F_o	average $\hat{se}_B$	0.489	0.319	
	SE of $\hat{se}_B$	(0.264)	(0.203)	
Case 2.	$\hat{\beta}_n$ average (SE)	1.015 (0.279)	1.007 (0.073)	1
discrete F_o	average $\hat{se}_B$	0.286	0.098	
	SE of $\hat{se}_B$	(0.067)	(0.013)	
Case 3.	$\hat{\beta}_n$ average (SE)	0.824 (1.355)	0.986 (0.391)	1
continuous F_o	average $\hat{se}_B$	0.430	0.275	
	SE of $\hat{se}_B$	(0.336)	(0.098)	

In Table 1, the entries in the column corresponding to n stand for the results with sample size n. The results in Table 1 suggest that $\hat{\beta}$ is consistent in all the three cases. Verify that $\frac{\partial E(\ln l(\beta))}{\partial \beta} \neq E(\frac{\partial \ln l(\beta)}{\partial \beta})$ in all the three cases, thus we cannot use $\hat{\sigma}^2_{\hat{\beta}}$ (see (3.5)) to estimate the variance of $\hat{\beta}$. In each simulation, using the bootstrap method described in Efron and Tibshirani (21) or Davison and Hinkley (20), we resampled (with replacement) B times. Each resample size is n. The bootstrap estimate of $\sigma_{\hat{\beta}}$, denoted by $\hat{se}_B$, is the sample standard error (SE) of the B estimates $\hat{\beta}$ based on the B resamples. As suggested by Efron and Tibshirani, we set $B = 30$. The entries corresponding to the row of "average $\hat{se}_B$" and the row of "SE of $\hat{se}_B$" are the sample means and the sample standard errors of these $\hat{se}_B$ in the 500 simulations, respectively. The differences between the sample SE's of $\hat{\beta}$ and the bootstrap estimates $\hat{se}_B$ are within 2 standard errors for the three cases, except for Case 3 when $n = 30$. This suggests that the bootstrap estimator $\hat{se}_B$ of the $\sigma_{\hat{\beta}}$ is appropriate.

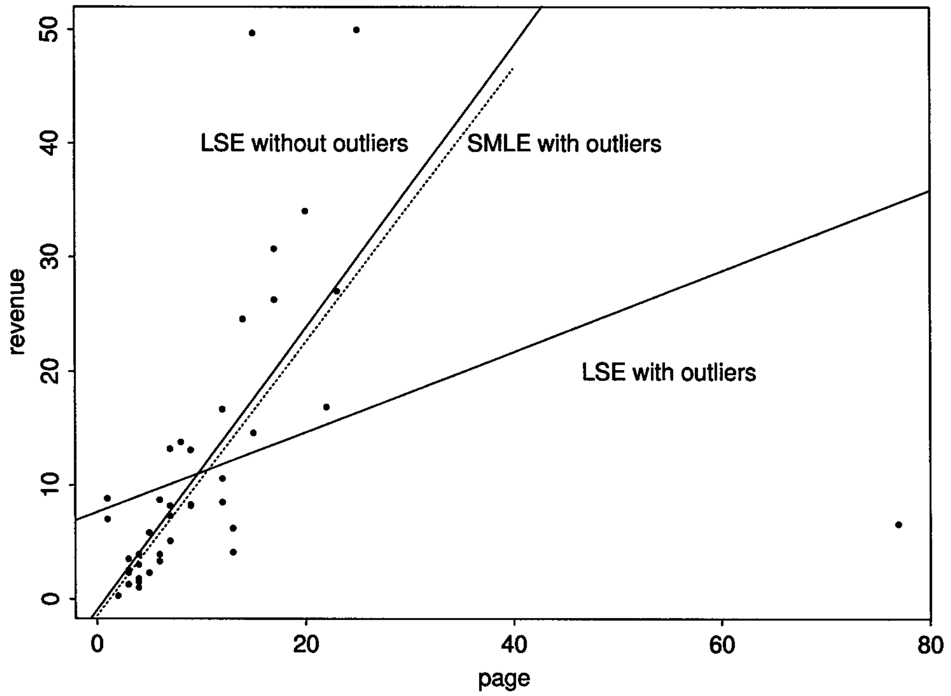
We also carried out simulation under the assumption that ϵ has a normal distribution or an exponential distribution. The results are similar and are not reported here.

5. APPLICATIONS

In this section, we apply our method to several data sets, including complete data ($L = R$ w.p.1), right-censored data and interval-censored data. We also compare the SMLE to the BJE, which is also an M-estimator (see Li and Zhang (12)).

Example 5.1. (Magazine advertising (Chatterjee and Price ((22), p. 257)). In a study of revenue from advertising, data were collected for 41 magazines in 1986. There was no censoring. Let X denote the number of pages of advertising and Y the advertising revenue.

Fig. 1. SMLE v.s. LSE



The 41 data are plotted in Figure 1. Roughly speaking, there are three outliers in the data set. They are (25, 50), (15, 49.7), (77, 6.6). The SMLE of β is unique for this data set. The SMLE and the LSE are significantly different (see the first block of Table 2). The entries in the second block of Table 2 are results after deleting the three outliers. From Table 2, it is seen that the SMLE of β does not change after deleting outliers, though the estimate of α changes.

In Figure 1, we also plot the fitted straight lines with and without deleting those three outliers. We further plot the fitted line by the SMLE method without deleting the outliers. From Figure 1, it is seen that the fitted line by the SMLE approach using the original data is very close to the least squares fitted line after deleting outliers. This suggests that the SMLE is robust while the LSE is not.

Table 2. Results on estimating (α, β)

		SMLE (SE)	LSE (SE)
with outliers	β	1.200 (0.196)	0.353 (0.1449)
	α	-1.427 (3.178)	7.604 (2.3466)
without outliers	β	1.200 (0.1379)	1.238 (0.138)
	α	-0.642 (1.410)	-0.962 (1.409)

Example 5.2. (Application to the Stanford heart transplant data). The data and detailed description can be found in Miller ((5), p. 156). In this data, right-censored survival time, indicator of death, indicator of rejection, 2 covariates including T5 mismatch and age of the recipient at time of transplant were recorded for 69 patients. For illustrative purposes, Buckley and James (4) compared their method to Miller's estimator and Cox method by fitting a simple linear regression for all death against age. We compare our method to the BJE using the same data under the log linear regression model.

Our algorithm results in an SMLE $\hat{\beta} = -9.77$ with a standard error (SE) 1.56. The SMLE is significantly negative and is consistent with our a priori guess, namely, younger patients fare better in the surgery.

There are 3 BJE's of β: -0.028, 0.014 and 0.016 with the SE 0.015 (only the first estimate and its SE were reported in Buckley and James (4)). They are all not significant at level 0.05. This means that there is no effect due to age.

In this example, the SMLE make more practical sense than the BJE, and thus is more reliable than the BJE.

Example 5.3. (Application to a cancer research data). Our data analysis is applied to a standard breast cancer relapse follow-up study based on data from 374 women with stages I - III unilateral invasive breast cancer surgically treated at Memorial Sloan-Kettering Cancer Center between 1985 and 1990. The median follow-up duration was 46 months. Relapse time was given by the time interval between surgery and the initial relapse. A relapse that took place between two successive follow-up visits was regarded as interval censored. If a patient did not relapse toward the end of the study, then her relapse time was right censored. Of the 374 observations, 300 were right censored (no relapse), 21 were left censored and 53 were strictly interval censored. Bone marrow micrometastasis (BMM) was determined for each woman at the time of surgery. An important question is whether remission duration is related to the extent of initial tumor burden defined as number of BMM cells detected. We compute the BJE and $\hat{\beta}$ under the log linear regression model.

One expects that the larger the BMM was, the shorter the patient survived. Thus $\beta < 0$. The BJE of β is -0.012 and our estimate is $\hat{\beta} = -0.059$ with a (bootstrap) standard error 0.02. Note that the BJE does not fall in the interval $(\hat{\beta} - 2SE, \hat{\beta} + 2SE)$, so they are significantly different. Moreover, the SMLE leans more toward the correct direction than the BJE.

Example 5.4. (Application to a breast cosmesis data). We applied our procedure to the interval-censored data set published in Finkelstein and Wolfe (23). The data is a result of a retrospective study to compare early breast cancer patients who have been treated with primary radiation therapy and a adjuvant chemotherapy to those treated with radiotherapy alone with respect to the cosmetic effects of their treatment. There is only one covariate, the group status, in this data set. There are 94 patients.

The BJE of β is -0.29 and our estimate $\hat{\beta}$ is -0.67 with a (bootstrap) standard error 0.336. The BJE falls in the interval $(\hat{\beta} - 2SE, \hat{\beta} + 2SE)$. Thus they are not significantly different.

6. CONCLUSION

In this paper, we propose algorithms for the SMLE of β. The procedure is actually applicable to complete data or right-censored data, as demonstrated in Examples 5.1 and 5.2.

We believe that each SMLE of β is consistent under interval censoring. The outline of the proof is as follows.

(1) By the definition of the SMLE and the strong law of large number, $\overline{\lim}_{n\to\infty} (\ln L(\hat{F}_{\hat{\beta}}, \hat{\beta}))/n \geq E(\ln L(F_o, \beta))/n$ a.s..

(2) By Fatou's Lemma, $\overline{\lim}_{n\to\infty} (\ln L(\hat{F}_{\hat{\beta}}, \hat{\beta}))/n \leq E(\ln L(F_*, b_*))/n$ a.s., where (F_*, b_*) is the limit of a convergent subsequence of the SMLE $(\hat{F}_{\hat{\beta}}, \hat{\beta})$.

(3) $E(\ln L(F, b))/n \leq E(\ln L(F_o, \beta))/n$ for each (F, b), and the equality holds only if $b = \beta$ by the Shannon-Kolmogorov inequality and the following assumption:

A3. $F \in \mathcal{F}$ and $\mathrm{P}\{F(W - bX) = F_o(W - \beta X) \text{ for } W = L \text{ or } R\} = 1$ imply $b = \beta$.

(The assumptions made in Section 4 satisfy A3.)

Statements (1) and (2) imply that $E(\ln L(F_*, b_*))/n = E(\ln L(F_o, \beta))/n$, thus statement (3) implies that $b_* = \beta$. Since b_* is an arbitrary limiting point of $\hat{\beta}$, $\hat{\beta}$ is consistent.

In general, the BJE is not efficient and is not robust, as the BJE reduces to the LSE in the case of complete data and the LSE is not efficient and is not robust (see Draper and Smith ((24), p. 342)). We apply both the BJE procedure and the SMLE to complete data, right-censored data and interval-censored data. The SMLE seems quite robust and always give reasonable estimates. The BJE may give unreasonable estimates.

For the semi-parametric set-up, there are several estimation procedures, *e.g.*, the BJE and the M-estimators. They are all obtained by iterative algorithms. The current procedure is also obtained by an iterative algorithm, unless the data are Case 1 interval-censored data. In the latter case, the SMLE can be obtained by a non-iterative algorithm, as the GMLE has closed-form solution (see Ayer *et al.* (25)). The reason is as follows. There are two steps in obtaining all SMLE's: (1) finding all the discontinuity points b_1, ..., b_m ($m \leq 4n^2$, see (3.4)), and (2) computing the GMLE $\hat{F}_b$ and comparing $l(b)$, $b = b_i$'s (see algorithm 3.1). Since the GMLE with Case 1 interval-censored data has a closed-form expression, all the SMLE's of β can be obtained in finitely many steps. However, with Case 1 interval-censored data, the BJE and the M-estimates cannot be obtained by a non-iterative algorithm.

We only discuss the simple linear regression model in this paper. The method can be extended to the multiple linear regression model. However, the computation can only performed on high-speed computers. Also, only extension of Algorithm 3.2 together with Steps (i) and (ii) is feasible due to the heavy computation cost.

BIGLIOGRAPHY

(1) Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, I. Proc. Kon. Ned. Akad. v. Wetensch. A 53 386-392.

(2) Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* 35 73-101.

(3) Bickel, P.J. (1982). On adaptive estimation. *Ann. Statist.* 10 647-671.

(4) Buckley, J and James, L. (1979). Linear regression with censored data. *Biometrika* 66 429-436.

(5) Miller, R.G. (1981). *Survival Analysis. Wiley.* N.Y.

(6) Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Ann. Statist.* 18 303-328.

(7) Odell, P.M., Anderson, K.M. and D'Agostino, R.B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics* 48 951-959.

(8) Self, S.G. and Grossman, E.A. (1986). Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers. *Biometrics* 42 521-530.

(9) Rabinowitz, D. Tsiatis, A. and Aragon, J. (1995). Regression with interval-censored data. *Biometrika* 82 501-513.

(10) Li, L.X. and Pu, Z.W. (1999). Regression models with arbitrarily interval-censored observations. *Comm. in Statist., Theory and Methods* 28 1547-1563.

(11) Zhang, C.H. and Li, X. (1996). Linear regression with doubly censored data. *Ann. Statist.* 24 2720-2743.

(12) Li, G. and Zhang, C.H. (1998). Linear regression with interval censored data. *Ann. Statist.* 26 1306-1327.

(13) Osborne M.R. (1985). *Finite algoriths in optimization and data analysis.* Wiley N.Y.

(14) Kiefer, J and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* 27 887-906.

(15) L. Ingber (1989). Very fast simulation re-annealing. *Mathematical and Computer Modelling* 12 967-973.

(16) Yu, Q.Q. and Wong, G.Y.C. (2003). The Semi-parametric MLE in linear regression with right-censored data. *Journal of Statistical Computation and Simulation*, 73 833-848.

(17) Cosslett, S.R. (1987). Efficient bounds for distribution-free estimator of the binary choice and the censored regression models. *Econometrica* 55 559-585.

(18) Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *JASA* 69 169-173.

(19) Wellner, J.A. and Zhan, Y. (1997). A hybrid algorithm for computation of the NPMLE from censored data. *JASA* **92** 945-959.

(20) Davison, A.C. and Hinkley, D.V. (1997) (1997). *Bootstrap methods and their application.* Cambridge Press. N.Y.

(21) Efron, B and Tibshirani, R.J. (1993). *An introduction to the bootstrap.* Chapman & Hall. N.Y.

(22) Chatterjee S. and Price B. (1991). *Regression analysis by example.* Wiley, N.Y.

(23) Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* 41 933-945.

(24) Draper, N.R. and Smith, H. (1981). *Applied Regression analysis. Wiley.* N.Y.

(25) Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T. and Silverman, E. (1955). An empirical distribution function for sampling incomplete information. *Ann. Math. Statist.* 26 641-647.

Reference (13)

Bone marrow micrometastasis is a significant predictor of long-term relapse-free survival for breast cancer by a non-proportional hazards model.

George YC Wong Ph D 1, Qiqing Yu Ph.D 1 and Michael P Osbome, MD 1.1 Preventive Oncology Research, Strang Cancer Prevention Center, New York, NY, United States, 10021 .

Background: Long-term predictive significance of the presence of bone marrow micrometastases (BMM) on breast cancer relapse is a substantively important question for clinicians. Two published long-term studies, the Royal Marsden study (Lancet 1999; 354:197-200) and our recent ASCO abstract (Proc Am Soc Clin Oncol 2002,21:228), have both concluded that BMM was not a significant predictor of relapse-free survival (RFS) by Cox proportional hazards (PH) regression analysis. However the RFS curves comparing presence and absence of BMM were separated in both these data sets. Diagnostic plots for PH assumption indicated that our RFS data were not consistent with such an assumption Consequently, Cox regression was inappropriate for the data and the logrank test was invalid. We analyzed our BMM data using a semi-parametric non-PH regression model.

Material and Methods: BMM was determined using monoclonal antibodies to cytokeratin at the time of initial surgery in 375 women with unilateral T1-2NO (56%), T1-2N1 (43%) and T3-4(1%) breast cancer. Relapse time was interval censored between two successive follow-up times. RFS data were analyzed using a regression model with a nonparametric error distribution that does not involve PH assumption. Statistical inference was based on a semi-parametric maximum likelihood estimation procedure.

Results: Median follow-up was 8 years (range 1 month-15 years). BMM was detected in 124 (35%) patients Contingency table analysis showed that BMM did not correlate with the standard prognostic variables of lymph node status (LN), tumor diameter (TD), estrogen and progesterone receptor levels. In the univariate non-PH RFS analysis, BMM was significant at p=0.05. In the multivariate analysis, BMM was still significant at p=0.05 in the presence of LN and TD. In contrast, only LN and TD were needed in the multivariate analysis by Cox regression.

Discussion: Recent published long-term studies using Cox regression have led to result that BMM is not a significant predictor for RFS. Using a novel non-PH regression model, we were able to demonstrate BMM as a statistically independent significant predictor of RFS in a multivariate regression model incorporating both LN and TD as covariates.